OUTCOME OF HARVEST SEASON

A Project Report

Submitted in the partial fulfillment of the requirements

for the award of the degree of

Bachelor of Technology in

Computer Science and Engineering

By

S.Juhitha-180030396

A.Chandana-180031031

under the supervision of

Dr. Chayan paul

PROFESSOR – Department of CSE

Koneru Lakshmaiah Education Foundation

(Deemed to be University estd., u/s 3 of UGC Act 1956)

Greenfields, Vaddeswaram, Guntur (Dist.), Andhra Pradesh - 522502

November, 2019

DECLARATION

The Project Report entitled “OutCome Of Harvest Season”is a record of bonafide work of **S.Juhitha(180030396)** and **A.Chandana(180031031)** submitted in partial fulfilment for the award of B.Tech in Computer Science and Engineering to the K L University. The results embodied in this report have not been copied from any other departments/University/Institute.

Sabbineni Juhitha-180030396

Aduri Chandana-180031031

CERTIFICATE

This is to certify that the Project Report entitled " **Outcome of Harvest Season**" is being submitted by **Sabbineni Juhitha** and **Aduri Chandana** submitted in partial fulfillment for the award of B.Tech in Computer Science and Engineering to the K L University is a record of bonafide work carried out under our guidance and supervision.

The results embodied in this report have not been copied from any otherndepartments/ University/Institute.

Signature of the Co-Supervisor

Signature of the Supervisor

Signature of the HOD

Signature of the External Examiner

ACKNOWLEDGEMENTS

It is with great pleasure to express our gratitude to our honourable President Sri**. Koneru Satya Narayana**, for giving the opportunity and platform with facilities in accomplishing the project based laboratory report.

We express sincere gratitude to HOD **Mr.V.Hari Kiran** for his leadership and constant motivation provided in successful completion of our academic semester. I record it as my privilege to deeply thank for providing us the efficient faculty and facilities to make our ideas into reality.

We express our sincere thanks to our project supervisor **Chayan paul** his novel association of ideas, encouragement, appreciation and intellectual zeal which motivated us to venture this project successfully.

Finally, it is pleased to acknowledge the indebtedness to all those who devoted themselves directly or indirectly to make this project report success.

S.Juhitha(180030396)

A.Chandana(180031031)

ABSTRACT

Recently Machine Learning concepts evolved in every sector even in agriculture sector where they are making effort to help the farmers to make their harvest season go smooth and have a healthy plantation at the end of the season. Well, however they are many other certain factor where they are yet to be developed in the agriculture sector and also have to evolve technologies which might help even better than before for farmers in their irrigation. Now, some of the factors that affect the crop are cyclone, pesticides, more number of insects in the respective farm etc.Now, the only thing that we can control is the amount of pesticides and if possible the number of insects. Here, I took a dataset with the details of harvest season crop plantation and the outcome of the crop at the end of that particular season. Based on the details and data points, I tried to know what amount of pesticides would be certainly good for the respective harvest season by the analysing the crop status i.e., whether the crop is alive ,or damaged by pesticides, or damaged by the some other reasons with the help of a ML algorithm that would fit the model better by doing comparative study.

TABLE OF CONTENTS

S.No.	Chapter	Title	PageNumber
1.	1	Introduction	1
	1.1	Objectives	2
	1.2	Background and Literature Survey	2
	1.3	Organization of the Report	3
2.	2	Outcome of Harvest Season	3
	2.1	Proposed System	4-9
	2.2	Working Methodology	10-13
	2.3	System Details	14
	2.3.1	Software	14
	2.3.2	Hardware	14
3.	3	Implementation	15-25
4.	4	Results and Discussion	26
5.	5	Conclusion and Future Works	27
6.	6	References	28

1- INTRODUCTION

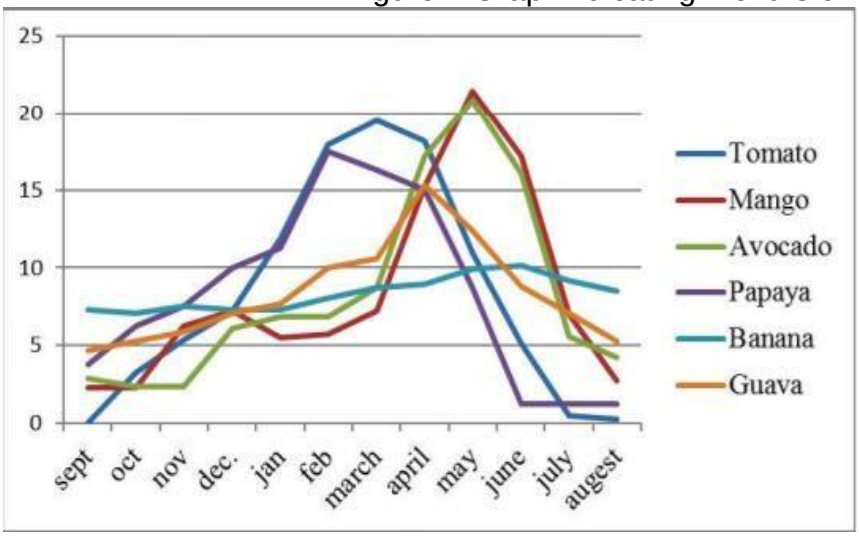
1.Introduction

Recently Machine Learning concepts evolved in every sector even in agriculture sector where they are making effort to help the farmers to make their harvest season go smooth and have a healthy plantation at the end of the season. Well, however they are many other certain factor where they are yet to be developed in the agriculture sector and also have to evolve technologies which might help even better than before for farmers in their irrigation.

Now, Pesticides are really good for crop while harvest period only when they are limited to certain measured amounts that required by the whole harvest season. If this is not the case then farmers would end up with the damage of whole crop in the farm and they would be falling into lots of debts and go through the tough times. But this would not be the situation if we predict the status of the crop in the early stages of harvesting so that farmer can take actions accordingly.

I took a dataset that contains the details of previous harvest crop plantation like the amount of pesticides, number of insects, soil type, crop type etc., and with respective to the crop damage status of particular harvest plantation. Here I had chosen the better fit algorithm to my model by doing comparative study by using multiple algorithms like Random Forest, KNN,Decision tree,Gaussian NB, Xgboost, lightgbm.

Figure1: Graphindicating months of high loss



1.1Objective

The following are the objectives of this project :

- To provide an efficient model to farmers for their best practices.
- To Enhance the Crop status after the whole harvest season
- To predict whether the crop is alive and or damaged by pesticides or damaged by some other reason.
- To provide the farmers a user friendly front end to use this model and use them for their crop prediction.

1.2 Background and Literature Survey

In this section, we review some of the significant works carried on in the agriculture stream for crop prediction.[1] Soil is the main source for agriculture, In this paper authors build a model of soil classification where they would suggest the crop based on soil serial accordingly. In this dataset, it

contains of nutrients that are required for the crop and their amount of quantity required for particular based on this they predict the soil quality for harvest season and crop plantation [2].

Also, there are some work done using data mining as it is the new technology that is rising in these days to forecast the agriculture productivity[3].This is a comparative model where authors worked to find the desired data model which is good in accuracy of forecasting productivity of yield.

1.3 Organization of the Report

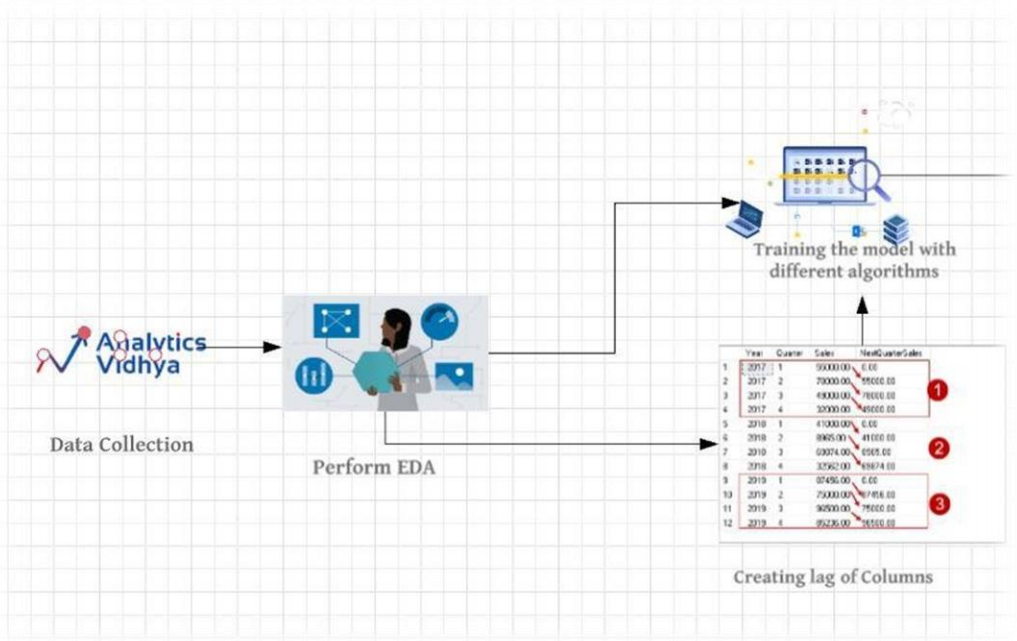
- Chapter 2 contains the proposed system, methodology, performance metrics, hardware and software details.
- Chapter 3 Implementation
- Chapter 4 Results And Discussions
- Chapter 5 concludes the report.
- Chapter 6 gives references.

CHAPTER 2

OUTCOME OF THE HARVEST SYSTEM

This Chapter describes the proposed system, working methodology, performance metrics, and software and hardware details.

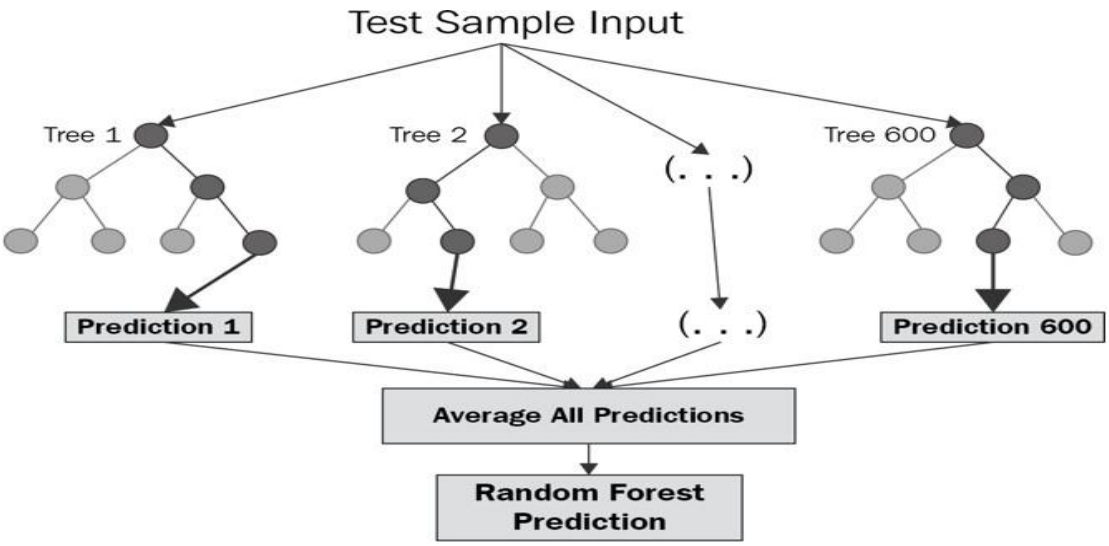
2.1 Proposed System



The following block diagram (figure 2) shows the system architecture of this project.

Figure 1: System Block Diagram

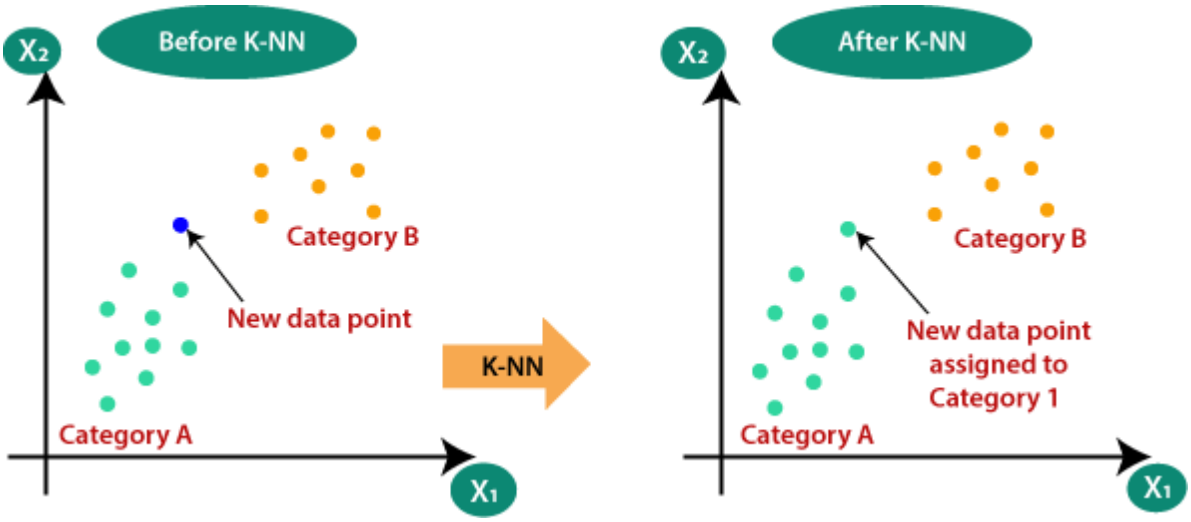
2.1.1Random Forest



As we all know, Random Forest is nothing but the collection of Decision Trees, later it chooses the predictions of all the decision trees and will do the average of all the predictions of decision tree classifier which result it for training the model more efficiently. Random forest requires all the input data points or variables to predict because it does the calculation with each every column with respect to the predictor or dependent variable and builds its own hierarchy which is more efficient for model fit.

2.1.2 K-Nearest Neighbour classifier

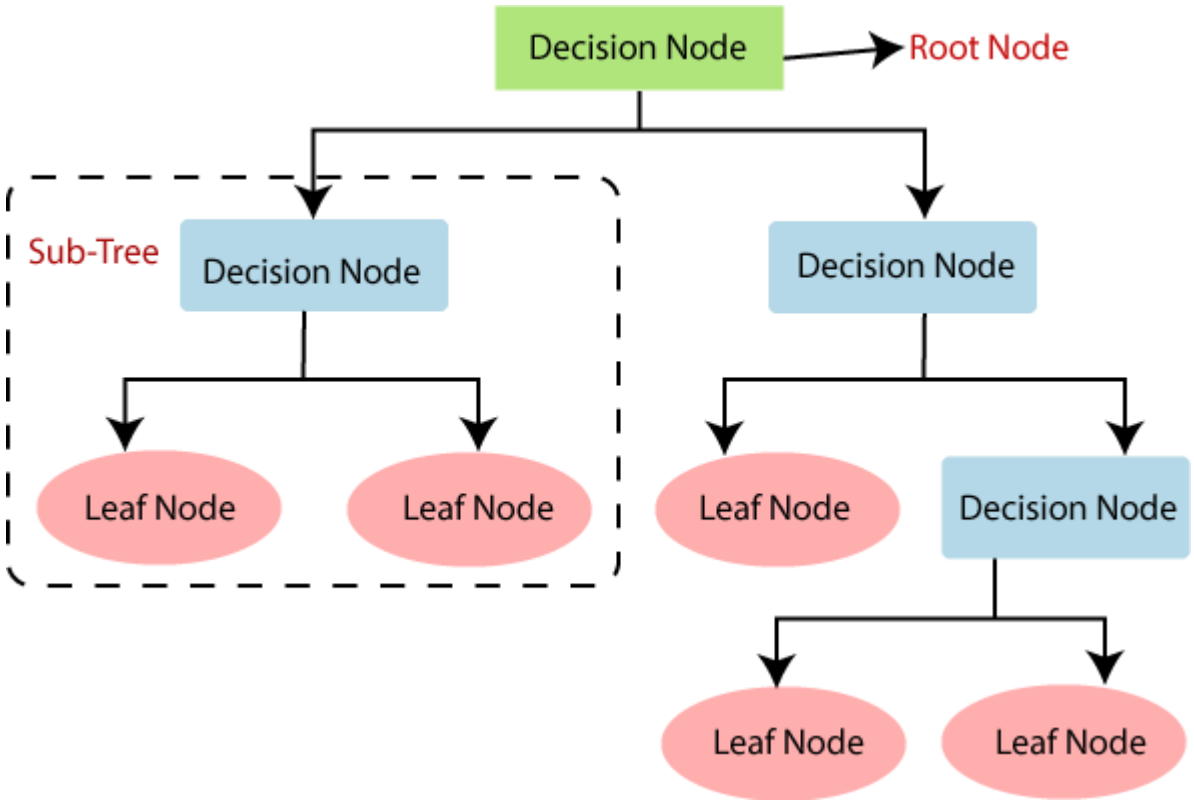
As we know K-Nearest Neighbor is a classification algorithm, it divides all the data points into certain categories with respective to their behavior and similarity between the data points later if you give a new point then it calculates the distance between the particular point with respective to the all categories present in there and with results of point which means the smallest distance that it got with respective to the categories it assigns the new point to that particular category.



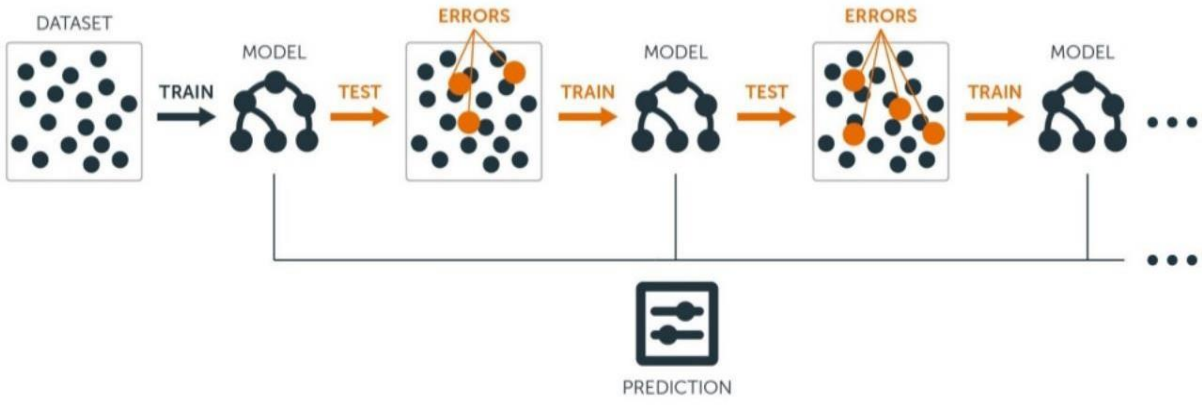
2.1.3 Decision Tree Classifier

we know, Decision Tree classifier is a classification algorithm which belonged to the supervised machine learning technique. It primarily chooses its root node for the later process then it chooses its first leaf node by calculating the gini-index of every column with respective to its roots node then the

highest value that we got of all those columns will considered as the first leaf node and then the process repeats with considering the following leaf node as the root node and builds a hierarchy.



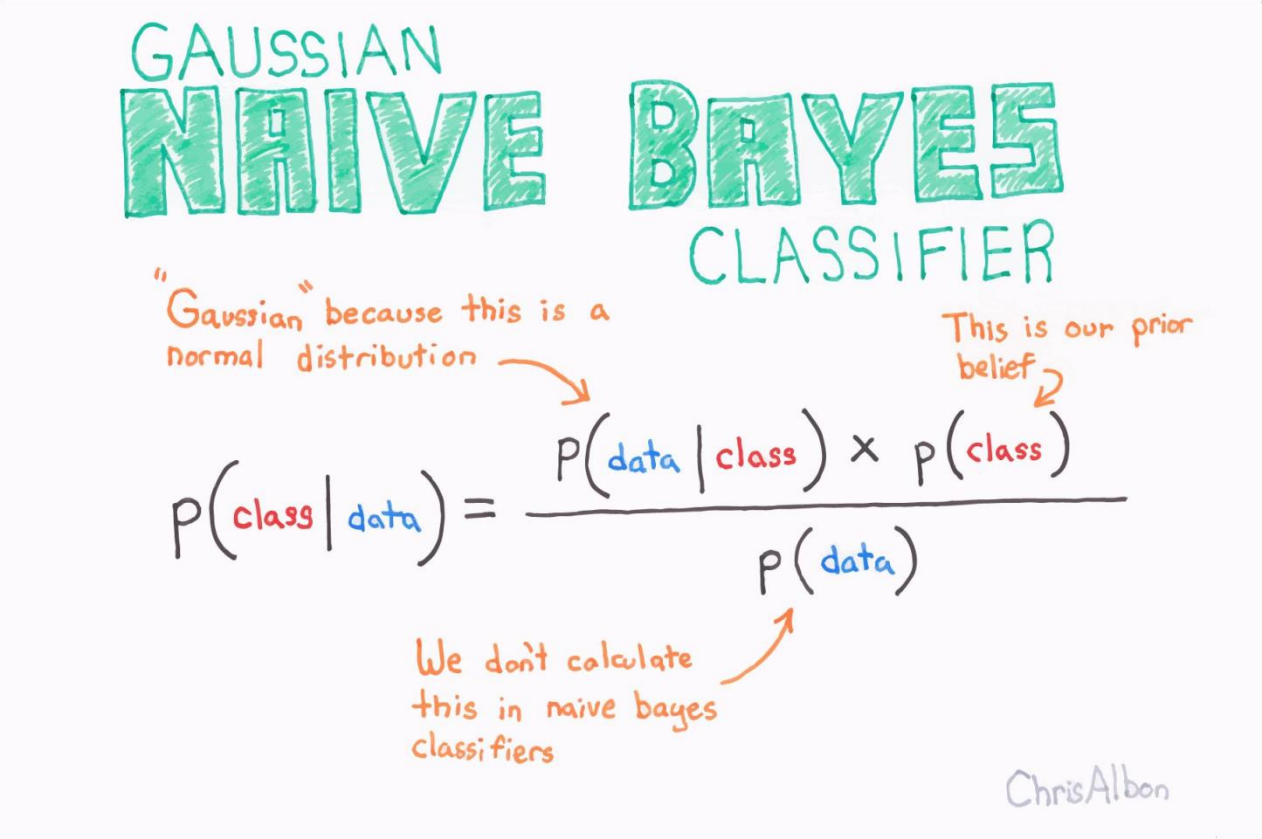
Gradient Decent



Gradient decent is a technique which calculates the loss of the model and decreases the error in the model. They are some algorithms which uses this technique to build an affiant model.

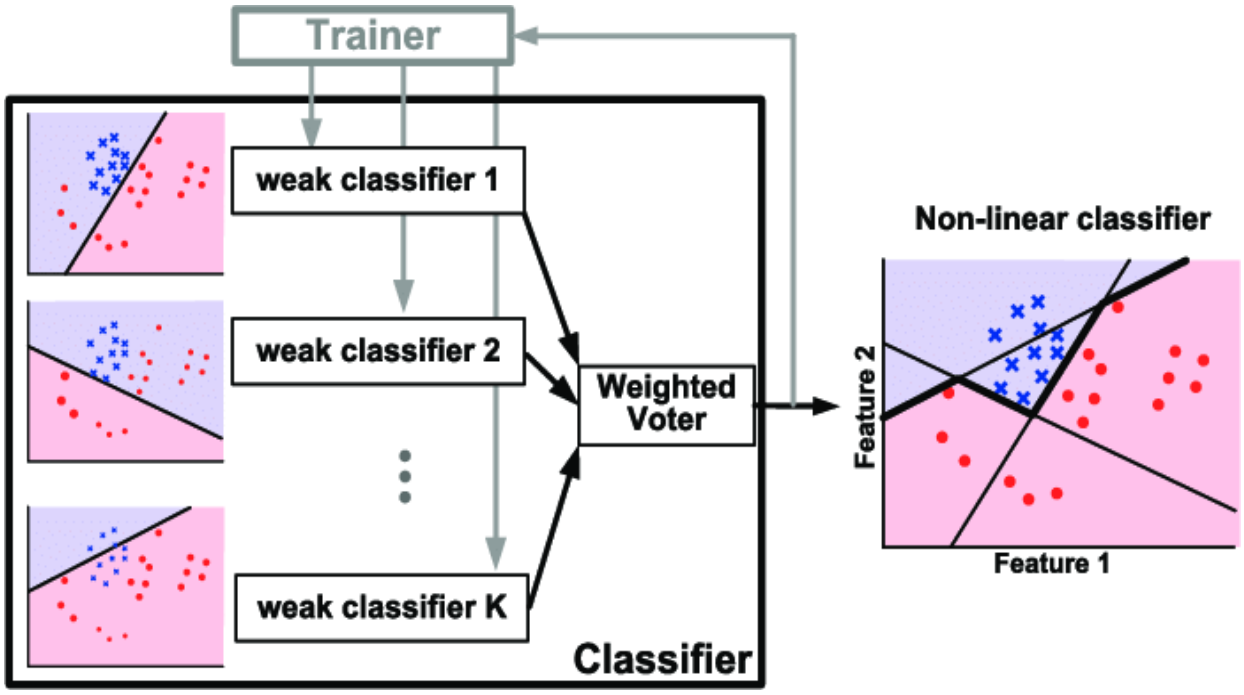
2.1.4 Gaussian NB

As we know that Gaussian NB is nothing but the extended version of Naïve bayes which means it builds on probability of the sample set that it is given but it also uses the Gaussian normal distribution to calculate the probability of the particular class with respective the predictor class or the dependent variable. So when it has a new point it calculates the probability of that particular point with the dependent variables or sample set and assign to the class variable which has high probability value.



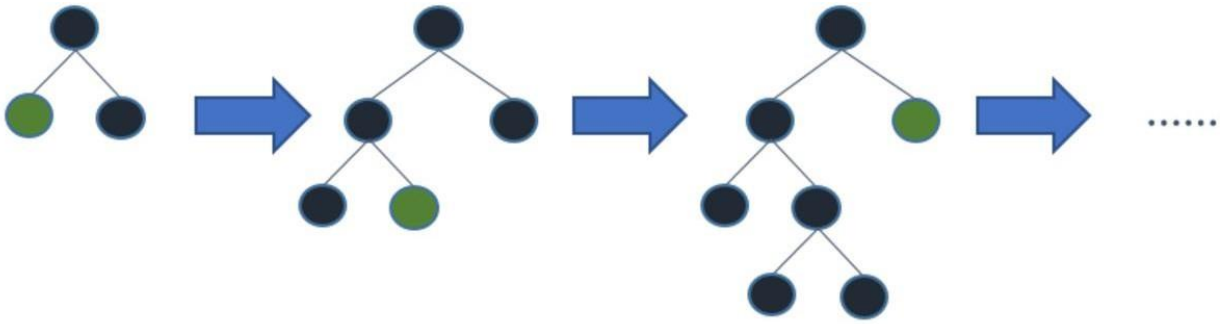
2.1.5 Ada Boost

As we know, Ada boost is a gradient decent Algorithm and it follows the same hierarchy as the decision tree. It calculates the loss from the week learners of the particular sample set and it will make the non-linear classifier to a linear classifier which builds a efficient model. It follows level wise hierarchy with respective to its root node and following leaf nodes.



2.1.6 LightGBM

As we know that LightGBM is also a gradient boosting Algorithm and it follows the same hierarchy as the decision tree to calculate its root node and then following nodes with respective to its root nodes. It calculates the loss from the week learners of the particular sample set and it will make the non-linear classifier to a linear classifier which builds a efficient model. It follows the leaf – wise growth makes it stand out of all algorithms and also it takes lesser time of execution when compared with other algorithms.



Leaf-wise tree growth

2.1.7 XGBoost

As we know that XGBoost is also a gradient boosting Algorithm and it follows the same hierarchy as the decision tree to calculate its root node and then following nodes with respective to its root nodes. It calculates the loss from the week learners of the particular sample set and it will make the non-linear classifier to a linear classifier which builds a efficient model. It follows the level – wise growth makes and it's okay with collinear and non-standardized features as well but it has less computational time compared with lightGBM.

2.2 Working Methodology

I took the data from a hack-a-thon conducted by Analytical vidya. There are two approaches that I did with the data points to bring out the better true positives out from the new data prediction points. Firstly, I did the whole inspection on the data where there are continues data points in the all columns there are 8 independent variables and a dependent variable (predictor) with 3 unique data points [0,1,2] i.e.,[Crop is alive, damaged by pesticides, damaged by some reason] and there are some missing values in the data. Now, I did the data visualization to get better understanding of the data and then I found that there are some outliers present in the data. So, I replaced them with the mean of those particular columns and then I certainly put my data into algorithms and observed the results and found out the better algorithm that fits to my model but with minimal amount of accuracy, precision, recall and F1-Score. Hence I did approach another way to increase my model accuracy. Here, I extended my columns with all possible ways of grouping my columns together and again I put this into algorithms where I observed that my accuracy increased compared with the last approach.

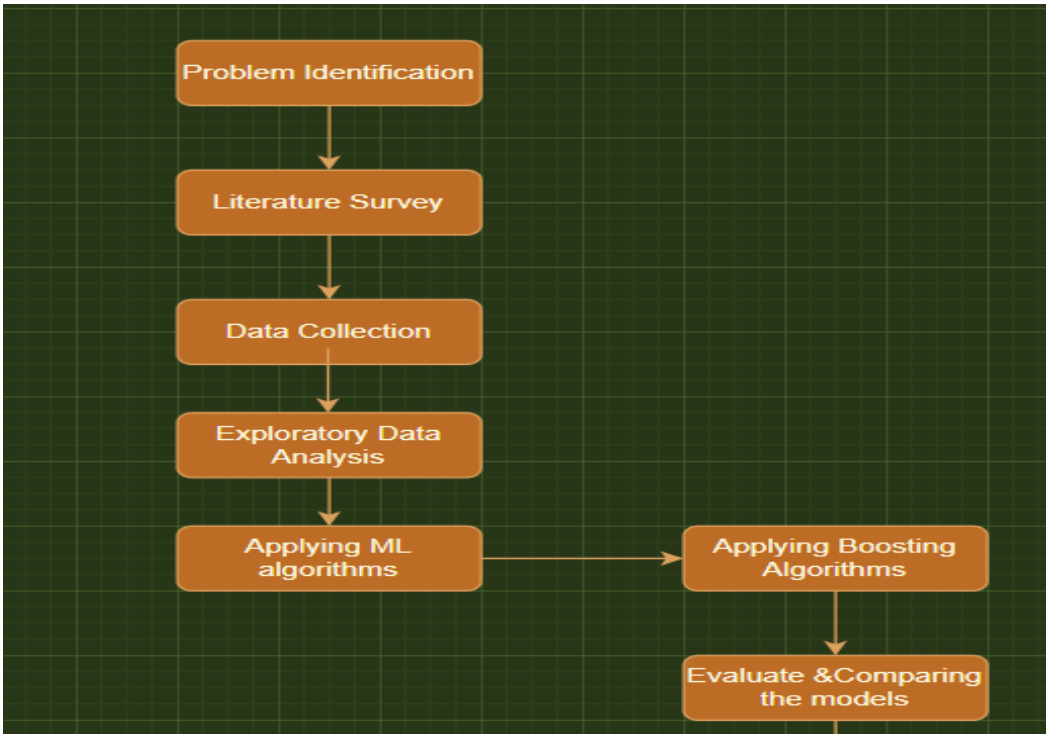


Figure 3: Flow Chart

2.2.1 Data set:

	A	B	C	D	E	F	G	H	I	J
1	ID	Estimated_Insects_Count	Crop_Type	Soil_Type	Pesticide_Use_Category	Number_Doses_Week	Number_Weeks_Used	Number_Weeks_Quit	Season	Crop_Damage
2	F00000001	188	1	0	1	0	0	0	1	0
3	F00000003	209	1	0	1	0	0	0	2	1
4	F00000004	257	1	0	1	0	0	0	2	1
5	F00000005	257	1	1	1	0	0	0	2	1
6	F00000006	342	1	0	1	0	0	0	2	1
7	F00000008	448	0	1	1	0		0	2	1
8	F00000009	448	0	1	1	0		0	2	1
9	F00000010	577	1	0	1	0	0	0	1	2
10	F00000012	731	0	0	1	0	0	0	2	0
11	F00000020	1132	1	0	1	0	0	0	1	2
12	F00000021	1212	1	0	1	0		0	3	0
13	F00000023	1575	0	0	1	0	0	0	1	1
14	F00000024	1575	0	1	1	0	0	0	2	1
15	F00000028	1575	1	1	1	0	0	0	2	1
16	F00000029	1575	1	1	1	0	0	0	2	2
17	F00000030	1785	1	1	1	0	0	0	2	1
18	F00000035	2138	0	1	1	0	0	0	1	1
19	F00000037	2401	0	1	1	0		0	1	1
20	F00000038	2401	1	1	1	0	0	0	2	1
21	F00000039	2401	1	1	1	0	0	0	2	1
22	F00000045	2999	0	1	1	0	0	0	3	1

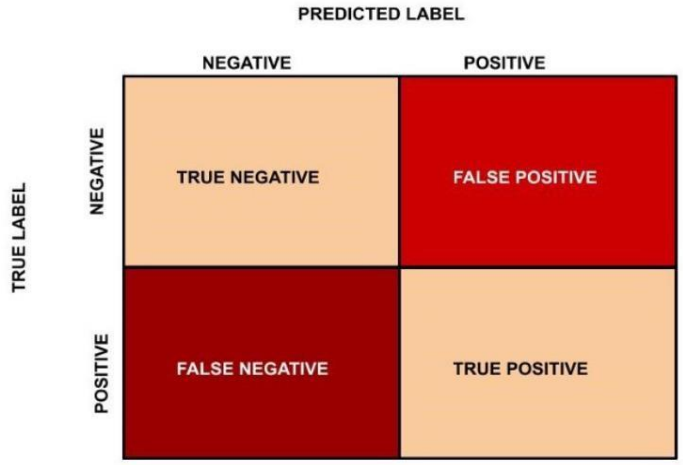
- ID – unique harvestID
- Estimated Insect Count – Number of insects estimated in particular HarvestSeason
- Crop Type – Type of the crop in theharvest
- Soil Type – Type of the Soil in theHarvest
- Pesticide Use Category – Category of the pesticides used in harvestseason
- Number Doses Week – No: of Doses with respective topesticides
- Number Weeks Used – No: of weeks that pesticides areused
- Number week Quit – Number of Weeks that the plantationquits
- Season – Type ofseason
- Crop Damage – (Dependent variable) shows the status ofcrop

2.2.2 Performancemetrics

- **Accuracy**

  The accuracy of a model is the number of new data points that the algorithm correctly classified. For instance, if the algorithm was tested on 100 new data points, and the algorithm correctly classified 97 of them — then we know that the accuracy is 97%.

  A confusion matrixis a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if we have an unequal number of observations in each class or if we have more than two classes in your dataset. Calculating a confusion matrix can give us a better idea of what the classification model is getting right and what types of errors it is making.



$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

- Precision

  The precision is the ratio tp/(tp+fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The best value is 1 and the worst value is0.

$$Precision = \frac{Correct\ Positive\ Predictions}{All\ Positive\ Prediction} = \frac{TP}{TP + FP}$$

- **Recall**

  The recall is the ratio tp/(tp+fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. The best value is 1 and the worst value is 0.

$$Recall = \frac{Correct\ Positive\ Predictions}{All\ Positives} = \frac{TP}{TP + FN}$$

- **F1-Score**

  F1-score is also known as balanced F-score or F-measure. It can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal.

$$F_1 = \left(\frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} \right) = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

- **Loss**

  The lower the **loss,** the better a model will be (unless the model has over-fitted to the training data). The loss is calculated on **training** and **validation** and its interpretation is how well the model is doing for these two sets. Unlike accuracy, loss is not a percentage. It is a summation of the errors made for each example in training or validation sets.

  Loss value implies how well or poorly a certain model behaves after every iteration of optimization. Ideally, one would expect the reduction of loss after each, or several, iteration(s).

2.3 SystemDetails

This section describes the software and hardware details of the system:

2.3.1 SoftwareDetails:-

OperatingSystem : Windows 10 or MAC(64bit)

Tools : Spyder or Jupyter notebook with required libraries, flaskframework

Libraries:

- **pandas** - Data manipulation and analysis
- **numpy** - Used to perform mathematical operations onarrays
- **nltk** - Contains NLP packages for stemming, Lemmatization, stopwordsetc
- **matplotlib** - Used for different plottings like line,scatter,barcharts,histograms
- **sklearn** - Contains different regression,classificationmodels,preprocessingmethods

2.3.2 HardwareDetails:-

Processor : Intel Core i5(7th Gen)

RAM : 4 GB ormore

HardDisk Drive : SATA

HDDCapacity : 1 TB

CHAPTER-3

IMPLEMENTATION

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
df_tn=pd.read_csv("train.csv")
df_tn["source"]="tarin"
df_tst=pd.read_csv("test.csv")
df_tst["source"]="test"
df=df_tn
df.head()
for i in df.columns:
    a=df[i].unique()
len(a)
    print(i,len(a))
    if len(a)<5:
        print(a)


df.nunique()
df.columns
df.isnull().sum()
df['Number_Weeks_Used'].fillna(df['Number_Weeks_Used'].mean(),inplace=True)
df.isnull().sum()
df.dtypes
sns.heatmap(df.corr(),annot=True,cmap='spring')
plt.figure(figsize=(12,5))
sns.catplot(x='Crop_Damage', data=df, palette="hls",kind='count',hue='Crop_Type')
```

```
plt.xlabel("Crop_Damage", fontsize=16)

plt.ylabel("Count", fontsize=16)

plt.title("Crop_Type Grouped Count", fontsize=20)

plt.xticks(rotation=45)

plt.show()

fig, [ax1,ax2,ax3] = plt.subplots(nrows=1,ncols=3,figsize=(15,5))

ax1=sns.countplot(x="Crop_Damage"
,hue="Pesticide_Use_Category",data=df[df["Crop_Damage"]==0],ax=ax1)

ax1.set_title("Crop Damage vs Insect Count for Crop Type")

ax2=sns.countplot(x="Crop_Damage"
,hue="Pesticide_Use_Category",data=df[df["Crop_Damage"]==1],ax=ax2)

ax2.set_title("Crop Damage vs Number Week Used")

ax3=sns.countplot(x="Crop_Damage"
,hue="Pesticide_Use_Category",data=df[df["Crop_Damage"]==2],ax=ax3)

ax3.set_title("Crop Damage vs Pesticide Use Category ")

plt.figure(figsize=(12,5))

g= sns.FacetGrid(df, col='Crop_Damage',size=5)

g = g.map(sns.distplot, "Number_Weeks_Used")

plt.show()

sns.barplot(x="Crop_Damage" ,y="Estimated_Insects_Count",hue="Crop_Type",data=df)

plt.figure(figsize=(12,5))

sns.catplot(x='Crop_Type',y='Number_Weeks_Used', data=df,
palette="hls",kind='bar',col='Crop_Damage')

plt.xticks(rotation=45)

plt.show()

df.describe()

df.drop(columns=["ID","source"],axis=1,inplace=True)

df.dtypes

df.plot(kind="box",subplots=True,layout=(5,5),figsize=(15,15))


df.loc[df['Number_Weeks_Used']>55,'Number_Weeks_Used'] =
np.mean(df["Number_Weeks_Used"])
```

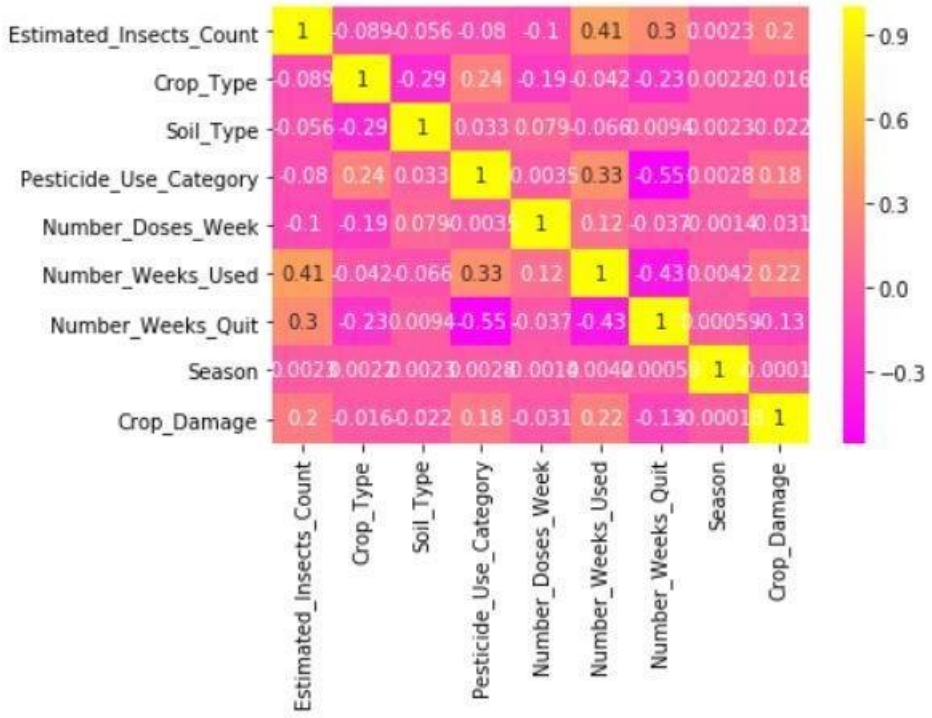
```
df.loc[df['Estimated_Insects_Count']>3500,'Estimated_Insects_Count'] =
np.mean(df["Estimated_Insects_Count"])

df.loc[df['Number_Weeks_Quit']>40,'Number_Weeks_Quit'] =
np.mean(df["Number_Weeks_Quit"])

df.loc[df['Number_Doses_Week']>80,'Number_Doses_Week'] =
np.mean(df["Number_Doses_Week"])


df.plot(kind="box",subplots=True,layout=(5,5),figsize=(15,15))

df.hist(figsize=(15,15), layout=(4,4), bins=20)

#Importing libraries

from sklearn.metrics import
accuracy_score,classification_report,confusion_matrix,roc_auc_score,roc_curve

from sklearn.linear_model import LogisticRegression

from sklearn.naive_bayes import MultinomialNB

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.svm import SVC

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import AdaBoostClassifier,GradientBoostingClassifier

from sklearn.model_selection import cross_val_score

from scipy.stats import zscore

from sklearn.preprocessing import LabelEncoder,StandardScaler

from sklearn.model_selection import train_test_split,GridSearchCV

from sklearn.decomposition import PCA

from sklearn.naive_bayes import GaussianNB

df_xc=df.drop(columns=['Crop_Damage'])

yc=df[["Crop_Damage"]]

from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

xc = sc.fit_transform(df_xc)

df_xc=pd.DataFrame(xc,columns=df_xc.columns)

#defining a function to find accuracy score, crossvalidation score for the given dataset
```

```
def max_acc_score(names,model_c,df_xc,yc):
accuracy_scr_max = 0
    for r_state in range(42,100):
train_xc,test_xc,train_yc,test_yc = train_test_split(df_xc,yc,random_state = r_state,test_size = 0.33,stratify = yc)
model_c.fit(train_xc,train_yc)
accuracy_scr = accuracy_score(test_yc,model_c.predict(test_xc))
        if accuracy_scr>accuracy_scr_max:
accuracy_scr_max=accuracy_scr
final_state = r_state
final_model = model_c
mean_acc = cross_val_score(final_model,df_xc,yc,cv=5,scoring="accuracy").mean()
std_dev = cross_val_score(final_model,df_xc,yc,cv=5,scoring="accuracy").std()
cross_val = cross_val_score(final_model,df_xc,yc,cv=5,scoring="accuracy")
print('\033[1m',"Results for model : ",names,'\n','\033[0m'
          "max accuracy score is" ,accuracy_scr_max ,'\n',
          "Mean accuracy score is : ",mean_acc,'\n',
          "Std deviation score is : ",std_dev,'\n',
          "Cross validation scores are :  " ,cross_val)
print(" "*100)
#Now by using multiple Algorithms we are calculating the best Algo which suit best for our data set
accuracy_scr_max = []
accuracy=[]
std_dev=[]
mean_acc=[]
cross_val=[]
models=[]
models.append(('Random Forest', RandomForestClassifier()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('Decision Tree Classifier', DecisionTreeClassifier()))
```

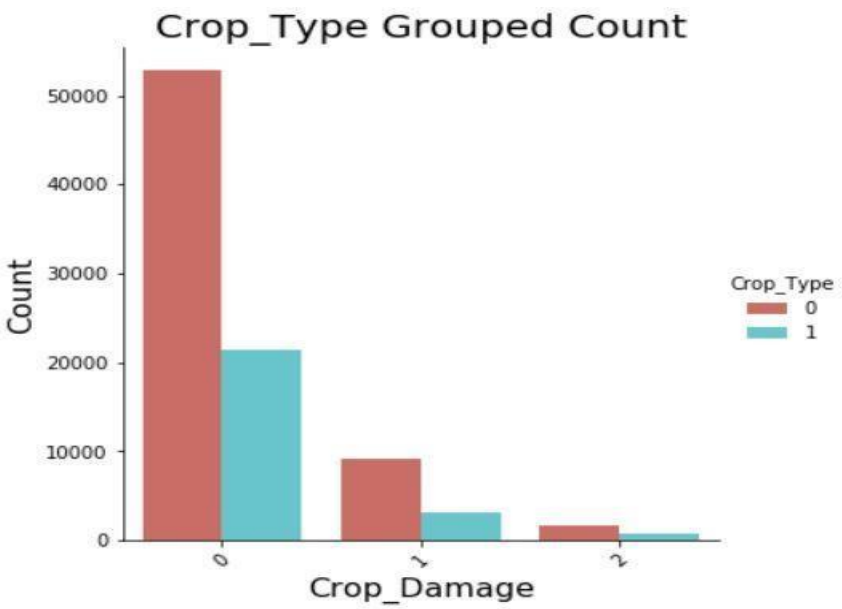
```
models.append(('Gaussian NB',GaussianNB()))



for names,model_c in models:
max_acc_score(names,model_c,df_xc,yc)
kNN=KNeighborsClassifier()
parameters={"n_neighbors":range(2,30)}
clf = GridSearchCV(kNN, parameters, cv=5,scoring="accuracy")
clf.fit(df_xc,yc)
clf.best_params_
#Again running KNeighborsClassifier with n_neighbor = 22
kNN=KNeighborsClassifier(n_neighbors=22)
max_acc_score("KNeighbors Classifier",kNN,df_xc,yc)
xc_train,xc_test,yc_train,yc_test=train_test_split(df_xc, yc,random_state =
80,test_size=0.20,stratify=yc)
kNN.fit(xc_train,yc_train)
yc_pred=kNN.predict(xc_test)
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import roc_auc_score

print("accuracy score is : ",accuracy_score(yc_test,yc_pred))
print("classification report \n",classification_report(yc_test,yc_pred))

cnf = confusion_matrix(yc_test,yc_pred)
sns.heatmap(cnf, annot=True, cmap = "Blues")

#Using adaboost classifier
from sklearn.ensemble import AdaBoostClassifier
ad=AdaBoostClassifier(n_estimators=10,learning_rate=1)
ad.fit(xc_train,yc_train)
```

```
ad_pred=ad.predict(xc_test)
print(accuracy_score(yc_test,ad_pred))
print(confusion_matrix(yc_test,ad_pred))
print(classification_report(yc_test,ad_pred))
import lightgbm as lgb
clf = lgb.LGBMClassifier(n_estimators=992,random_state=42,max_depth=
18,learning_rate=0.4,class_weight= {0: 0.44, 1: 0.4, 2: 0.37},min_data_in_leaf = 55,subsample =
0.7,objective='multiclass',reg_alpha = 1.7,reg_lambda = 1.11,colsample_bytree=0.7)
clf.fit(xc_train, yc_train)

lg_pred=clf.predict(xc_test)
print(accuracy_score(yc_test,lg_pred))
print(confusion_matrix(yc_test,lg_pred))
print(classification_report(yc_test,lg_pred))



import xgboost as xgb
model=xgb.XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
importance_type='gain', interaction_constraints='',
learning_rate=0.4, max_delta_step=0, max_depth=18,
min_child_weight=20, missing=0, monotone_constraints='()',
n_estimators=3000, n_jobs=10, num_parallel_tree=15,
      objective='multiclass', random_state=42, reg_alpha=1.7,
reg_lambda=1.11, scale_pos_weight=None, subsample=0.7,
tree_method='exact', use_label_encoder=True, validate_parameters=1,
      verbosity=None)
model.fit(xc_train, yc_train)
xgb_pred=model.predict(xc_test)
print(accuracy_score(yc_test,xgb_pred))
print(confusion_matrix(yc_test,xgb_pred))
```

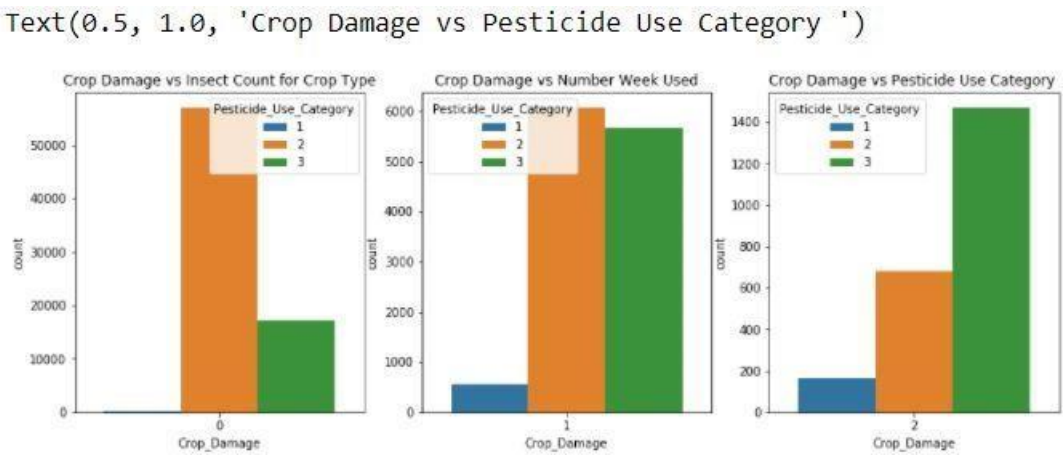
print(classification_report(yc_test,xgb_pred))



a. Correlation states the strong and weak bond between the columns respectively. So that it help us while we are filling the missing values, to do feature engineeringetc.

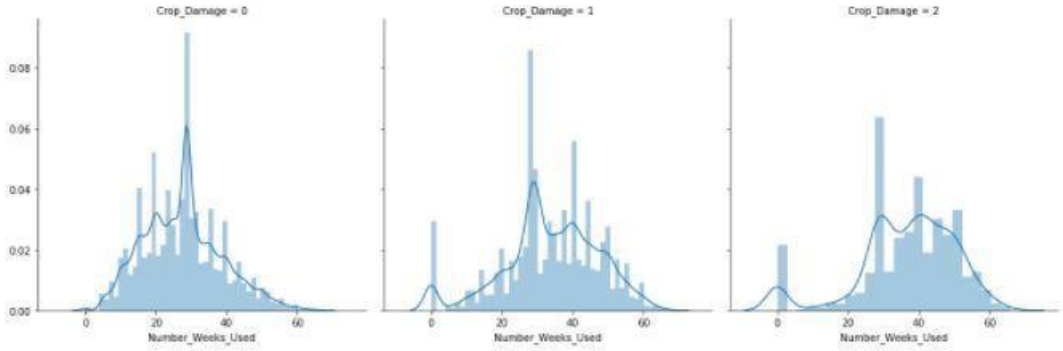


b. Here is the graph of crop_type Grouped count which is observed between the crop_type with respective to the crop damage where we got some insights like crop_type 0 is damaged more compared with crop_type1.



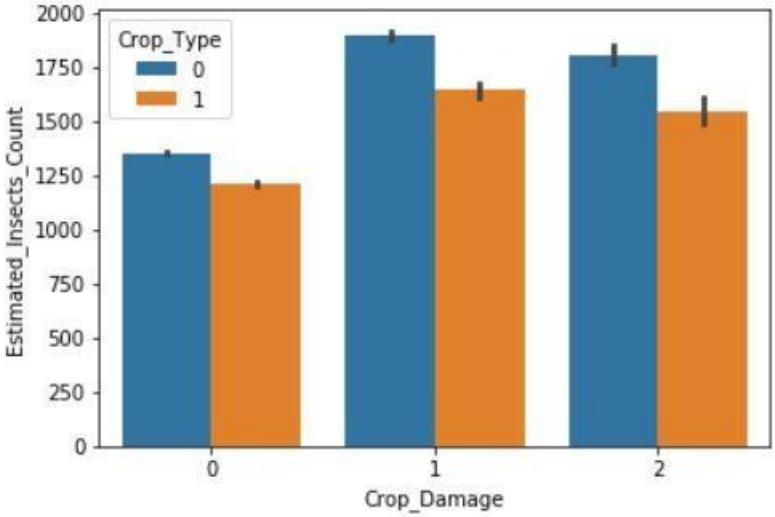
c. The above bar graph shows us the crop damage vs Pesticide usecategory.

```
<Figure size 864x360 with 0 Axes>
```



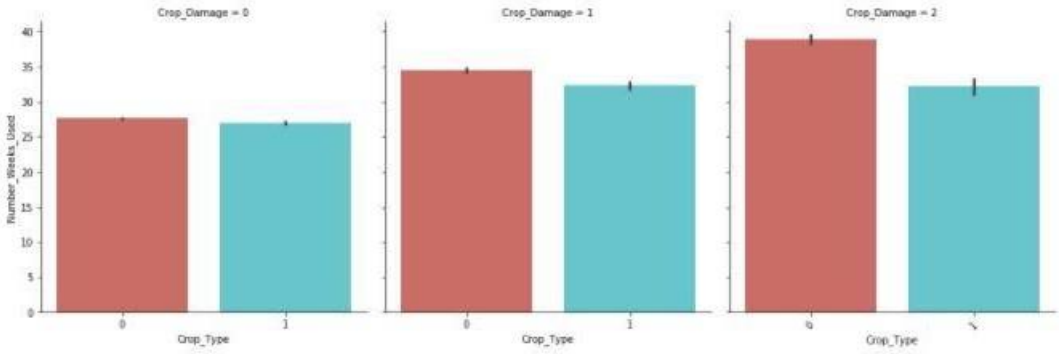
d. The above graph shows the distribution of Number weeks used with respective to the dependent classvariable.

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c8e08bc320>
```

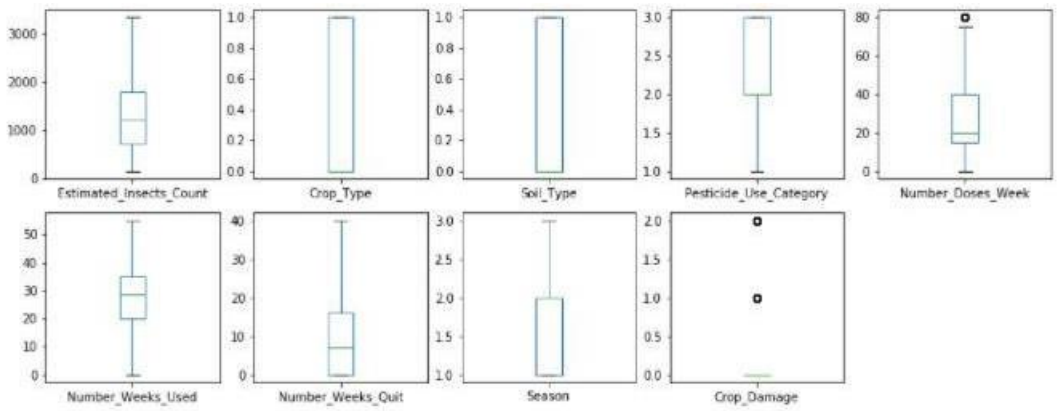


e. The above bar graph shows the crop damage with respective to the crop type and Estimated insectscount.

<Figure size 864x360 with 0 Axes>



f. The above boxplot shows us the outlier present in thedata.



g. The above boxplot shows us the outlier after the cleaning of data where the data is filtered with the mean of their respectivecolumns.

CHAPTER 4

RESULTS AND DISCUSSIONS

Outcome of harvest season is all about predicting the outcome or status of the crop plantation at the end of harvest season. Hence, I solved this problem using machine learning where I did a comparative study by following some algorithms and tried to fit my model in them they are., Random Forest Classifier, K-Nearest Neighbor, Decision Tree Classifier, Gaussian NB, Ada-Boost, LightGBM, XgBoost and later I observed the performance metrics of each algorithm by that I concluded LightGBM has high performance when compared with remaining algorithms.

Also, this is the approach where I performed exploratory data analysis, and trained my model then I only got 84.6% as the highest accuracy .

S.NO	ALGORITHM	ACCURACY	Precision	Recall	F1-Score
1	Random Forest Classifier	82.2%	0.74	0.81	0.73
2	K-Nearest Neighbours	84.3%	0.79	0.84	0.80
3	Decision Tree Classifier	75.3%	0.66	0.71	0.78
4	Gaussian NB	82.4%	0.75	0.80	0.74
5	Adaboost	84%	0.79	0.84	0.77
6	**Lightgbm**	**84.6%**	**0.82**	**0.85**	**0.81**
7	Xgboost	80%	0.72	0.70	0.65

CHAPTER 5

CONCLUSION AND FUTURE WORK

Outcome of harvest season is all about predicting the outcome or status of the crop plantation at the end of harvest season. This would give an idea for the farmers that how much amount of pesticides he has to use and can also take care of the insects in his field so that he can increase efficiency of the harvest crop to be healthy. I have the dataset with 8-independent and 1- dependent variable. Here, in the dependent variable I have 3 class labels where 0 - Crop is alive, 1– Damaged by pesticides. 2 – Damaged by some other reason.

By doing data inspection, exploratory data analysis and model building, I got a final accuracy of 84.6% .

In the future, I will try to research on some more ways to protect crop like controlling the over flow of water in field by measuring the certain amount that required by the fields. I will also do the research on chance to decrease the floods in the crop field by this the farmer ending up on debts will decreases and could live his life smoothly only then we could live our lives smoothly because farmers are the backbone of ourcountry.

REFERENCES

[1] Sk Al Zaminur Rahman, S.M. Mohidul Islam, Kaushik Chandra Mitra,‖ Soil Classification using Machine Learning Methods and Crop Suggestion Based on Soil Series‖,2018 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December, 2018.

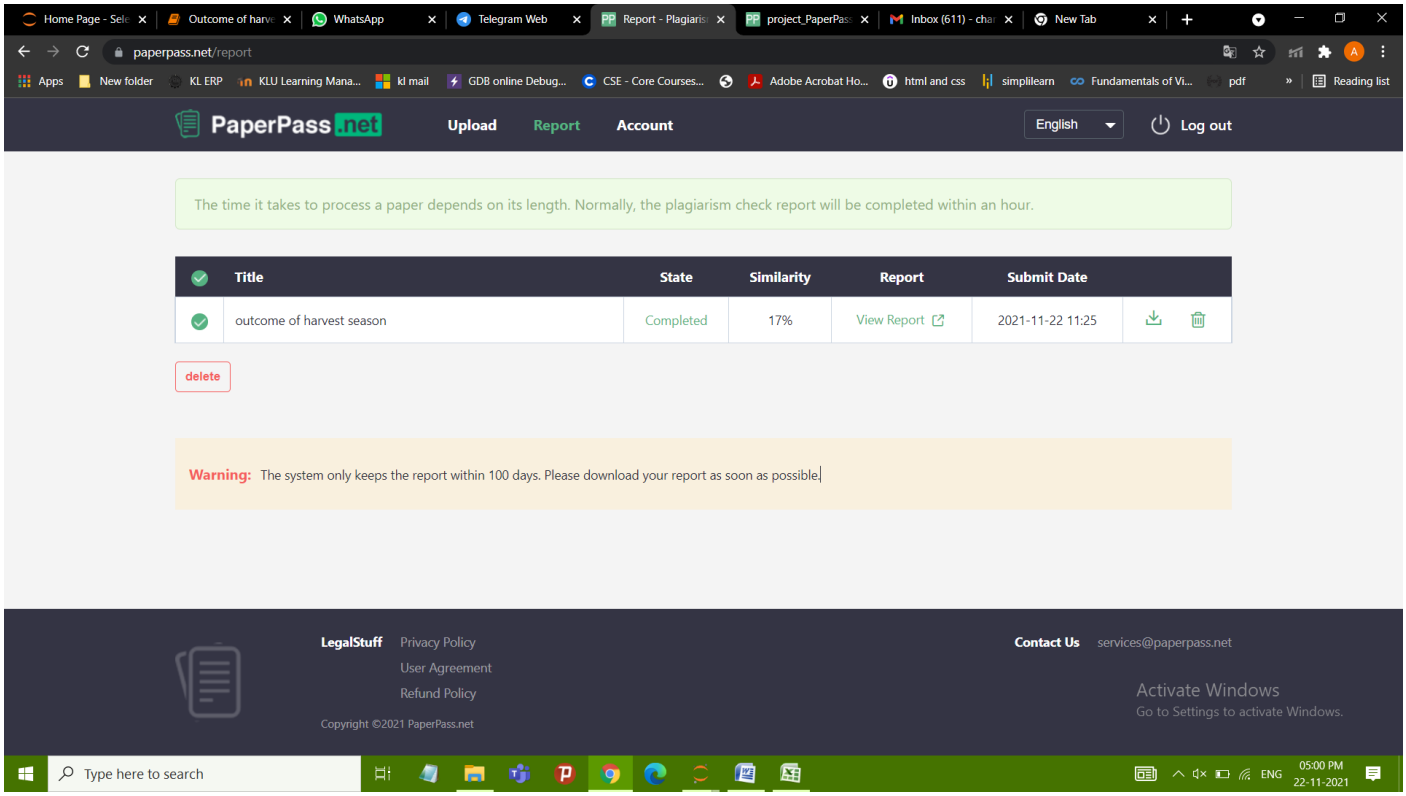
[2] S. Panchamurthi. M.E.,M.D.Perarulalan,A. Syed Hameeduddin,P. Yuvaraj,‖Soil Analysis and Prediction of Suitable Crop for Agriculture using Machine Learning‖, International Journal for Research in Applied Science & Engineering Technology(IJRASET),ISSN:2321-9653;IC Value:45.98;SJ Impact Factor:6.887,Volume 7 Issue III,Mar 2019.

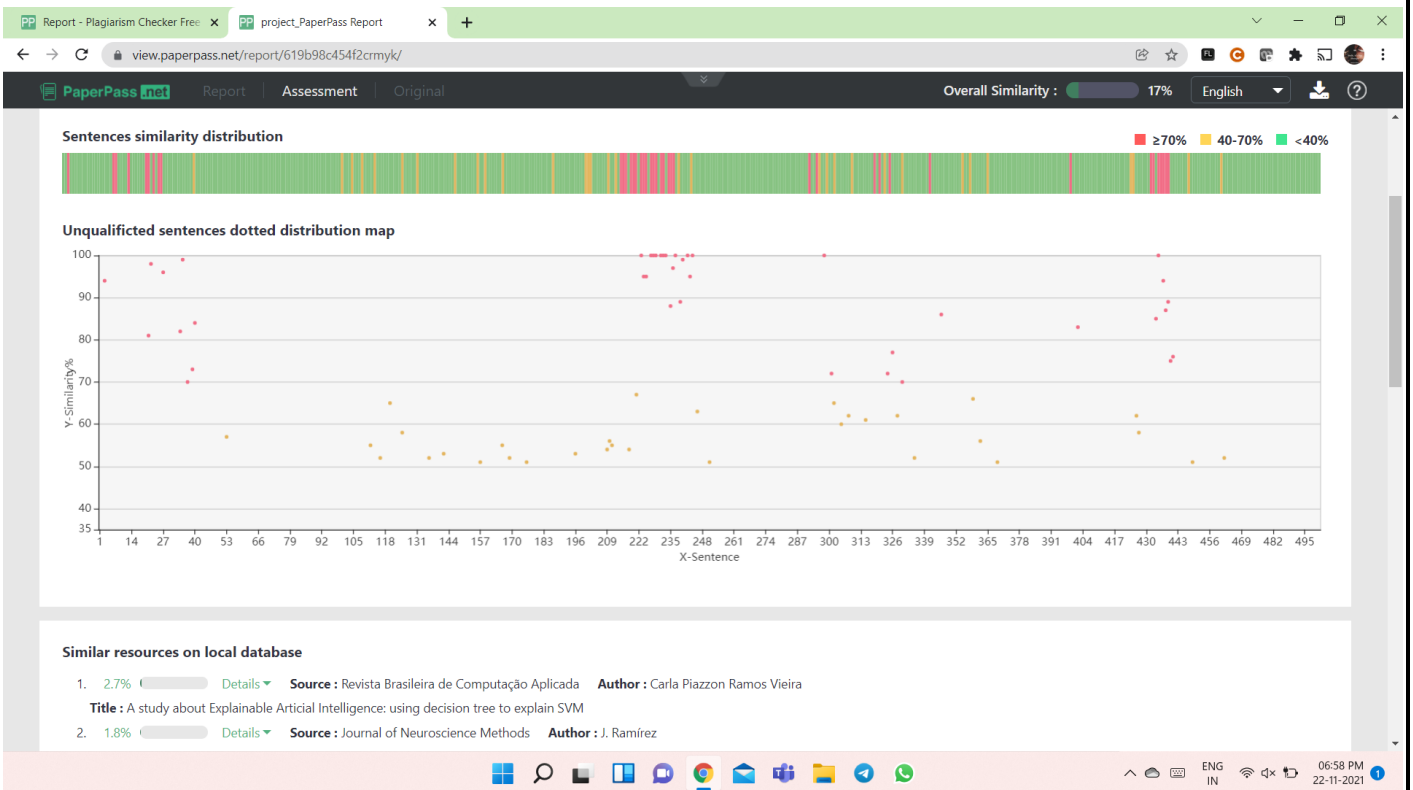
[3] D Ramesh,B Vishnu Vardhan,―Data Mining Techniques and Applications to Agricultural Yield Data,‖ International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.

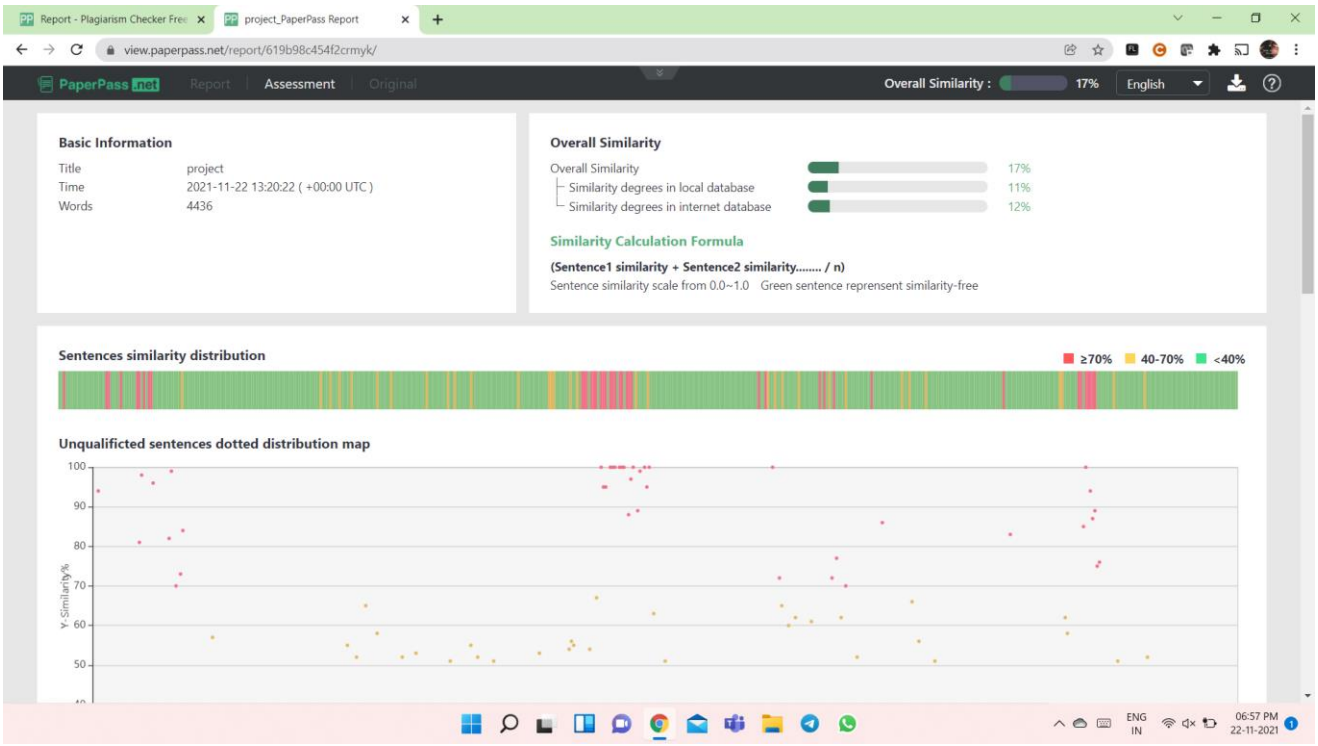
[4]https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.ht ml

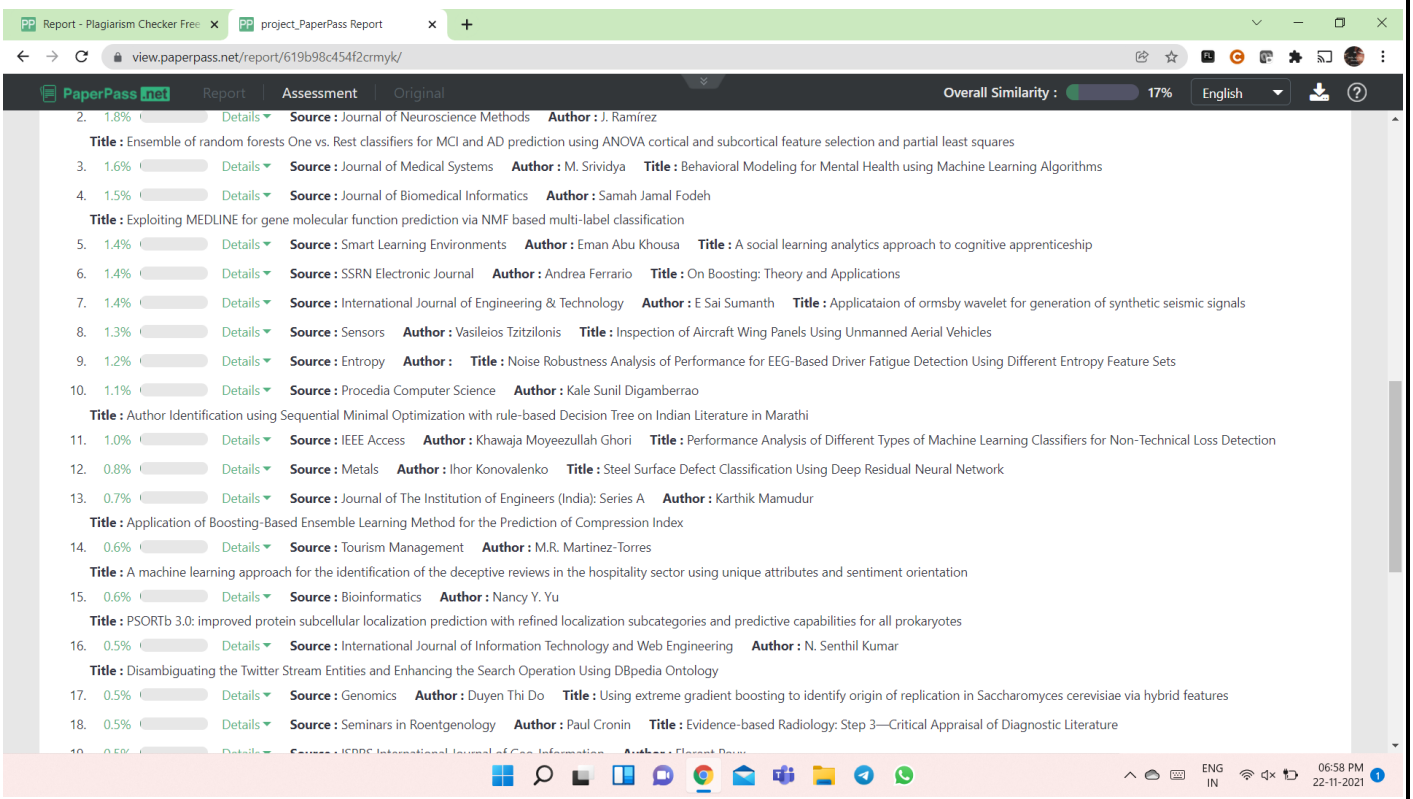
[5] https://machinelearningmastery.com/boosting-and-adaboost-for-machinelearning/
[6]https://scikitlearn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
[7]https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[8]https://xgboost.readthedocs.io/en/latest/

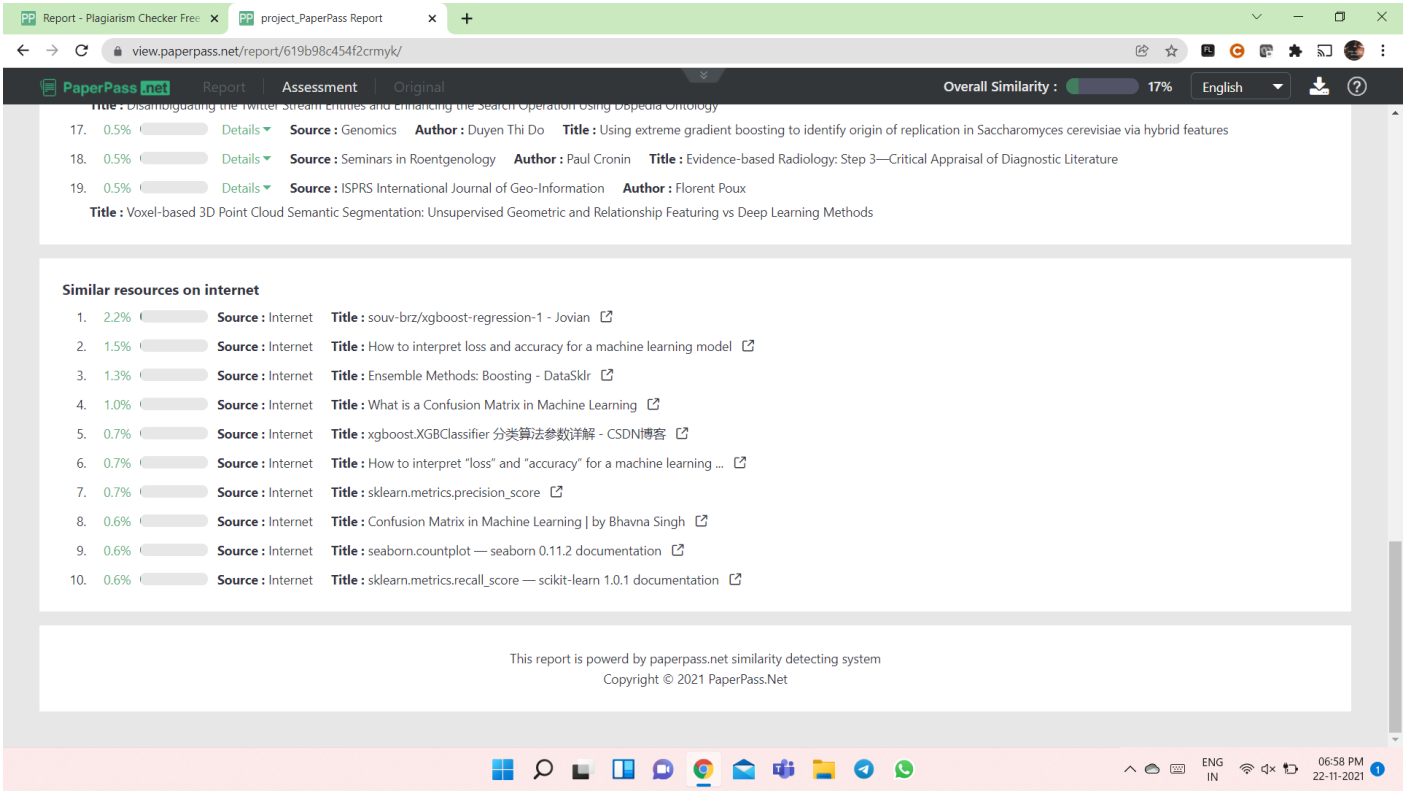
PLAGIARISM REPORT






Report - Plagiarism Checker Free
project_PaperPass Report
view.paperpass.net/report/619b98c454f2crmyk/
PaperPass .net
Report
Assessment
Original
Overall Similarity : 17%
English
Basic Information
Title project
Time 2021-11-22 13:20:22 (+00:00 UTC)
Words 4436
Overall Similarity
Overall Similarity 17%
Similarity degrees in local database 11%
Similarity degrees in internet database 12%
Similarity Calculation Formula
(Sentence1 similarity + Sentence2 similarity..... / n)
Sentence similarity scale from 0.0~1.0 Green sentence reprensent similarity-free
Sentences similarity distribution
≥70% 40-70% <40%
Unqualificted sentences dotted distribution map
Y-Similarity%
100
90
80
70
60
50
ENG
IN
06:57 PM
22-11-2021


Report - Plagiarism Checker Free
project_PaperPass Report
view.paperpass.net/report/619b98c454f2crmyk/
PaperPass .net
Report
Assessment
Original
Overall Similarity : 17%
English
2. 1.8% Details Source : Journal of Neuroscience Methods Author : J. Ramírez
Title : Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares
3. 1.6% Details Source : Journal of Medical Systems Author : M. Srividya Title : Behavioral Modeling for Mental Health using Machine Learning Algorithms
4. 1.5% Details Source : Journal of Biomedical Informatics Author : Samah Jamal Fodeh
Title : Exploiting MEDLINE for gene molecular function prediction via NMF based multi-label classification
5. 1.4% Details Source : Smart Learning Environments Author : Eman Abu Khousa Title : A social learning analytics approach to cognitive apprenticeship
6. 1.4% Details Source : SSRN Electronic Journal Author : Andrea Ferrario Title : On Boosting: Theory and Applications
7. 1.4% Details Source : International Journal of Engineering & Technology Author : E Sai Sumanth Title : Applicataion of ormsby wavelet for generation of synthetic seismic signals
8. 1.3% Details Source : Sensors Author : Vasileios Tzitzilonis Title : Inspection of Aircraft Wing Panels Using Unmanned Aerial Vehicles
9. 1.2% Details Source : Entropy Author : Title : Noise Robustness Analysis of Performance for EEG-Based Driver Fatigue Detection Using Different Entropy Feature Sets
10. 1.1% Details Source : Procedia Computer Science Author : Kale Sunil Digamberrao
Title : Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi
11. 1.0% Details Source : IEEE Access Author : Khawaja Moyeezullah Ghori Title : Performance Analysis of Different Types of Machine Learning Classifiers for Non-Technical Loss Detection
12. 0.8% Details Source : Metals Author : Ihor Konovalenko Title : Steel Surface Defect Classification Using Deep Residual Neural Network
13. 0.7% Details Source : Journal of The Institution of Engineers (India): Series A Author : Karthik Mamudur
Title : Application of Boosting-Based Ensemble Learning Method for the Prediction of Compression Index
14. 0.6% Details Source : Tourism Management Author : M.R. Martinez-Torres
Title : A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation
15. 0.6% Details Source : Bioinformatics Author : Nancy Y. Yu
Title : PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes
16. 0.5% Details Source : International Journal of Information Technology and Web Engineering Author : N. Senthil Kumar
Title : Disambiguating the Twitter Stream Entities and Enhancing the Search Operation Using DBpedia Ontology
17. 0.5% Details Source : Genomics Author : Duyen Thi Do Title : Using extreme gradient boosting to identify origin of replication in Saccharomyces cerevisiae via hybrid features
18. 0.5% Details Source : Seminars in Roentgenology Author : Paul Cronin Title : Evidence-based Radiology: Step 3—Critical Appraisal of Diagnostic Literature
ENG
IN
06:58 PM
22-11-2021


Report - Plagiarism Checker Free
project_PaperPass Report
view.paperpass.net/report/619b98c454f2crmyk/
PaperPass .net
Report
Assessment
Original
Overall Similarity : 17%
English
17. 0.5% Details Source : Genomics Author : Duyen Thi Do Title : Using extreme gradient boosting to identify origin of replication in Saccharomyces cerevisiae via hybrid features
18. 0.5% Details Source : Seminars in Roentgenology Author : Paul Cronin Title : Evidence-based Radiology: Step 3—Critical Appraisal of Diagnostic Literature
19. 0.5% Details Source : ISPRS International Journal of Geo-Information Author : Florent Poux
Title : Voxel-based 3D Point Cloud Semantic Segmentation: Unsupervised Geometric and Relationship Featuring vs Deep Learning Methods
Similar resources on internet
1. 2.2% Source : Internet Title : souv-brz/xgboost-regression-1 - Jovian
2. 1.5% Source : Internet Title : How to interpret loss and accuracy for a machine learning model
3. 1.3% Source : Internet Title : Ensemble Methods: Boosting - DataSklr
4. 1.0% Source : Internet Title : What is a Confusion Matrix in Machine Learning
5. 0.7% Source : Internet Title : xgboost.XGBClassifier 分类算法参数详解 - CSDN博客
6. 0.7% Source : Internet Title : How to interpret "loss" and "accuracy" for a machine learning ...
7. 0.7% Source : Internet Title : sklearn.metrics.precision_score
8. 0.6% Source : Internet Title : Confusion Matrix in Machine Learning | by Bhavna Singh
9. 0.6% Source : Internet Title : seaborn.countplot — seaborn 0.11.2 documentation
10. 0.6% Source : Internet Title : sklearn.metrics.recall_score — scikit-learn 1.0.1 documentation
This report is powerd by paperpass.net similarity detecting system
Copyright © 2021 PaperPass.Net
ENG
IN
06:58 PM
22-11-2021