WELLSFARGO CAMPUS ANALYTICS CHALLENGE

OPEN-PROJECT 2 – STAGE 5

APPLIED MACHINE LEARNING

JOANNA AUGUSTINE – 800656114 **JITHIN JACOB BENJAMIN JACOB - 800681973**

1) OBJECTIVE: Our project consists of three phases. In phase one we obtained the WellsFargoCampusAnalytics Challenge Data which had two spreadsheets namely Individuals and Resources Data where the various ways in which carbon emission was given.

Next we handled the data by removing NaN values and did some preprocessing on the data and made sure that the data was shaped accordingly by not creating fake data and by converting the datatypes to float64.

Thirdly we Conducted various Clustering methodologies and used different Classifiers on the Individuals and Resources data and did validation on it. We observed the accuracy and the Sum Squared error.

2) ANALYSIS ON CLUSTERING FOR RESOURCES DATA:

a) Clustering was conducted on the Resources reshaping and the Individual data Reshaping.Firstly we present the results for the Resources Reshaping method. The KMeans Algorithm was run on the "Resources" data.

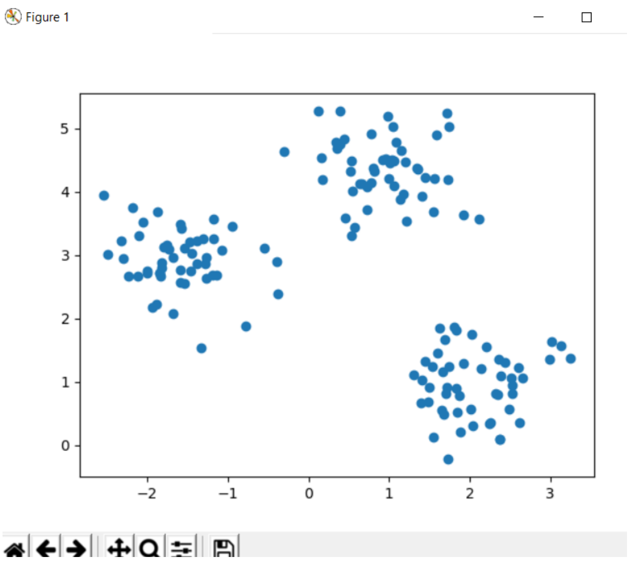
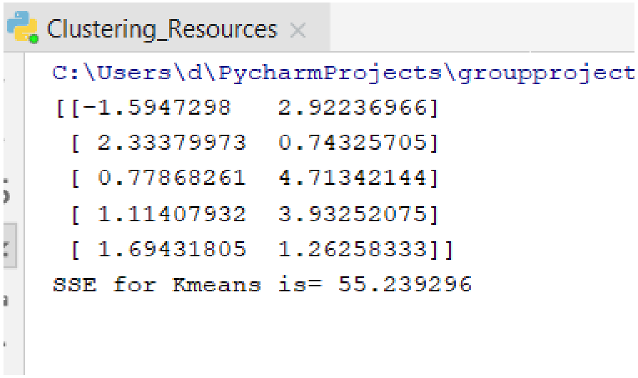


Fig 2.1 The KMeans Algorithm clustering on the Resources data



```
C:\Users\d\PycharmProjects\groupproject
[[-1.5947298   2.92236966]
 [ 2.33379973  0.74325705]
 [ 0.77868261  4.71342144]
 [ 1.11407932  3.93252075]
 [ 1.69431805  1.26258333]]
SSE for Kmeans is= 55.239296
```

Fig 2.2 The Sum Squared Error was found to be 55.2392

b) Next the same was done for hierarchial clusters using Agglomerative clustering

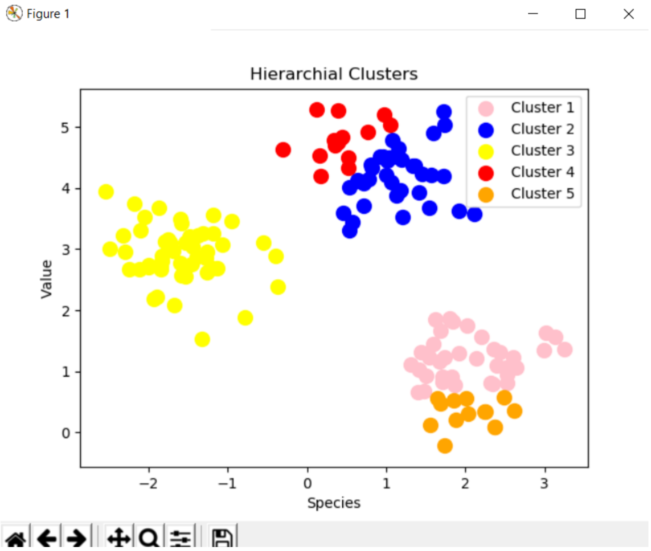
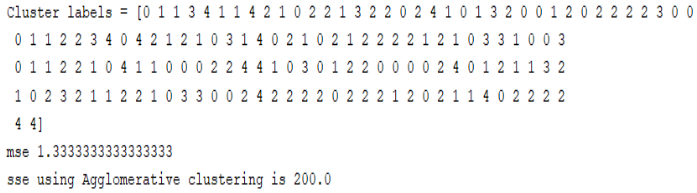


Fig 2.3 Clustering for Resources using Agglomerative Clustering



```
Cluster labels = [0 1 1 3 4 1 1 4 2 1 0 2 2 1 3 2 2 0 2 4 1 0 1 3 2 0 0 1 2 0 2 2 2 2 3 0 0
 0 1 1 2 2 3 4 0 4 2 1 2 1 0 3 1 4 0 2 1 0 2 1 2 2 2 2 1 2 1 0 3 3 1 0 0 3
 0 1 1 2 2 1 0 4 1 1 0 0 0 2 2 4 4 1 0 3 0 1 2 2 0 0 0 0 2 4 0 1 2 1 1 3 2
 1 0 2 3 2 1 1 2 2 1 0 3 3 0 0 2 4 2 2 2 2 0 2 2 2 1 2 0 2 1 1 4 0 2 2 2 2
 4 4]
mse 1.3333333333333333
sse using Agglomerative clustering is 200.0
```

Fig 2.4 The mean squared error and Sum squared error is outputted

c) Next is the DBScan clustering which was implemented

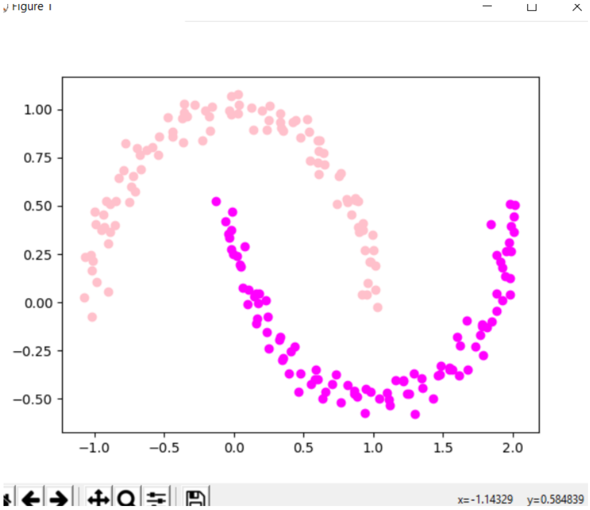
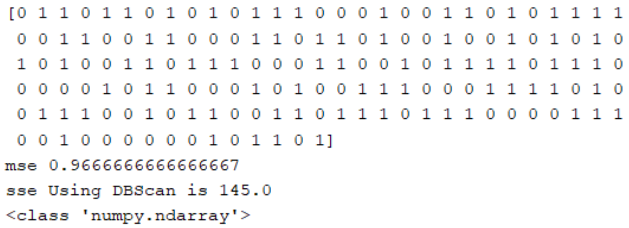


Fig 2.5 The moon shaped clusters were observed



```
[0 1 1 0 1 1 0 1 0 1 0 1 1 1 0 0 0 1 0 0 1 1 0 1 0 1 1 1 1
 0 0 1 1 0 0 1 1 0 0 0 1 1 0 1 1 0 1 0 0 1 0 0 1 0 1 0 1 0
 1 0 1 0 0 1 1 0 1 1 1 0 0 0 1 1 0 0 1 0 1 1 1 1 0 1 1 1 0
 0 0 0 0 1 0 1 1 0 0 0 1 0 1 0 0 1 1 1 0 0 0 1 1 1 1 0 1 0
 0 1 1 1 0 0 1 0 1 1 0 0 1 1 0 1 1 1 0 1 1 1 0 0 0 0 1 1 1
 0 0 1 0 0 0 0 0 0 1 0 1 1 0 1]
mse 0.9666666666666667
sse Using DBScan is 145.0
<class 'numpy.ndarray'>
```

Fig 2.6 The mean squared error and the sse was given

d) The same was conducted using the elbow method and the respective graph with the number of clusters and the WCSS is generated.

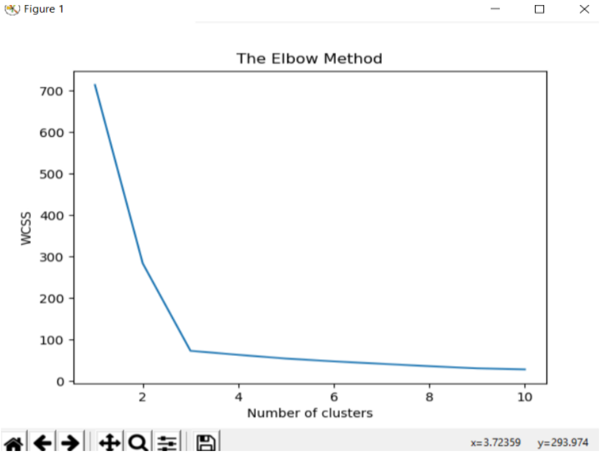
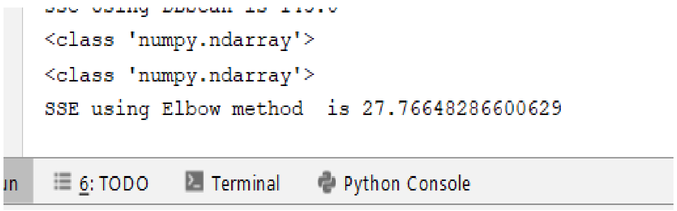


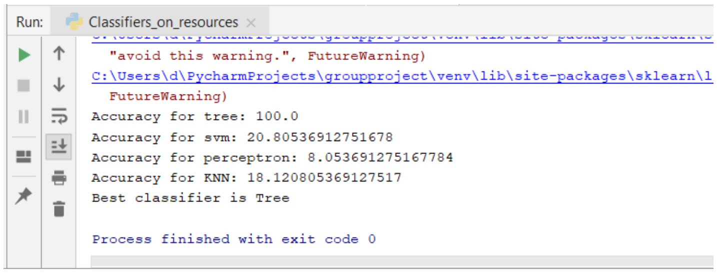
Fig 2.7 The WCSS vs the number of clusters was given as the output parameters



```
SSE using DBScan is 145.0
<class 'numpy.ndarray'>
<class 'numpy.ndarray'>
SSE using Elbow method  is 27.76648286600629
```

Fig 2.8 The SSE using elbow method was carefully observed.

3)ANALYSIS FOR DIFFERENT CLASSIFIERS ON RESOURCES DATA:



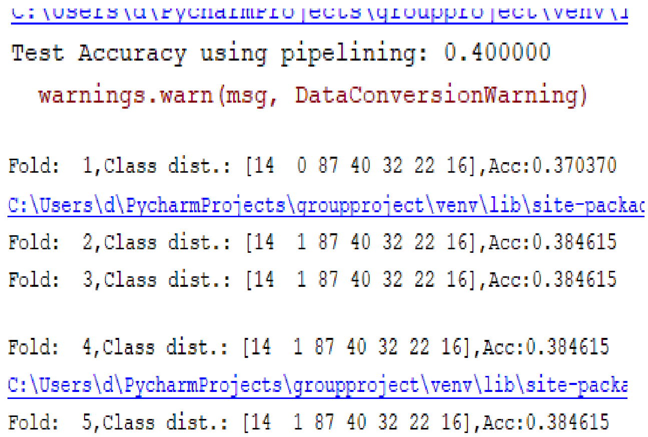
```
  "avoid this warning.", FutureWarning)
C:\Users\d\PycharmProjects\groupproject\venv\lib\site-packages\sklearn\l
  FutureWarning)
Accuracy for tree: 100.0
Accuracy for svm: 20.80536912751678
Accuracy for perceptron: 8.053691275167784
Accuracy for KNN: 18.120805369127517
Best classifier is Tree

Process finished with exit code 0
```

Fig 3.1 The best classifier was identified as the Tree classifier based on the accuracy metrics.

4)VALIDATION ON RESOURCES DATA:

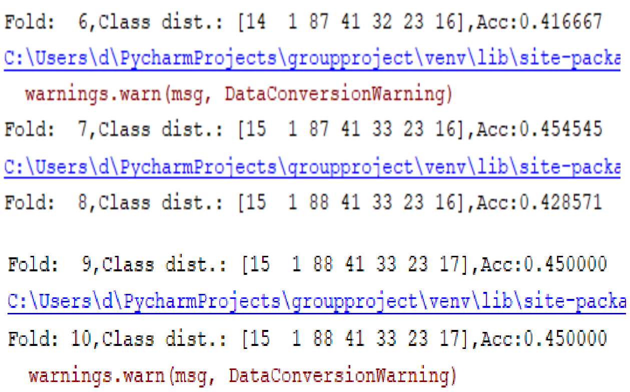
a) Then validation was done on the Resources data using the cross evaluation metrics such as pipelining



```
C:\Users\d\PycharmProjects\groupproject\venv\l
Test Accuracy using pipelining: 0.400000
  warnings.warn(msg, DataConversionWarning)

Fold:  1,Class dist.: [14  0 87 40 32 22 16],Acc:0.370370
C:\Users\d\PycharmProjects\groupproject\venv\lib\site-packag
Fold:  2,Class dist.: [14  1 87 40 32 22 16],Acc:0.384615
Fold:  3,Class dist.: [14  1 87 40 32 22 16],Acc:0.384615

Fold:  4,Class dist.: [14  1 87 40 32 22 16],Acc:0.384615
C:\Users\d\PycharmProjects\groupproject\venv\lib\site-packa
Fold:  5,Class dist.: [14  1 87 40 32 22 16],Acc:0.384615
```



```
Fold:  6,Class dist.: [14  1 87 41 32 23 16],Acc:0.416667
C:\Users\d\PycharmProjects\groupproject\venv\lib\site-packa
  warnings.warn(msg, DataConversionWarning)
Fold:  7,Class dist.: [15  1 87 41 33 23 16],Acc:0.454545
C:\Users\d\PycharmProjects\groupproject\venv\lib\site-packa
Fold:  8,Class dist.: [15  1 88 41 33 23 16],Acc:0.428571

Fold:  9,Class dist.: [15  1 88 41 33 23 17],Acc:0.450000
C:\Users\d\PycharmProjects\groupproject\venv\lib\site-packa
Fold: 10,Class dist.: [15  1 88 41 33 23 17],Acc:0.450000
  warnings.warn(msg, DataConversionWarning)
```

Fig 4.2 This was the validation outputs

5)CLUSTERING ON INDIVIDUAL DATA:

The same clustering methods was conducted using the individuals data and the accuracy and Sum Squared Errors were calculated.

a) Hierarchial clustering

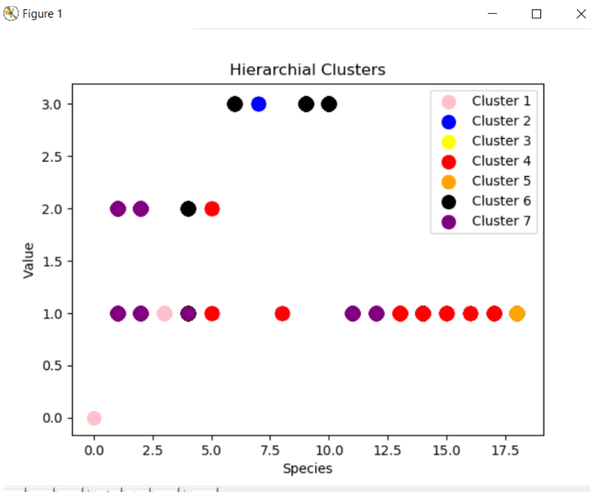
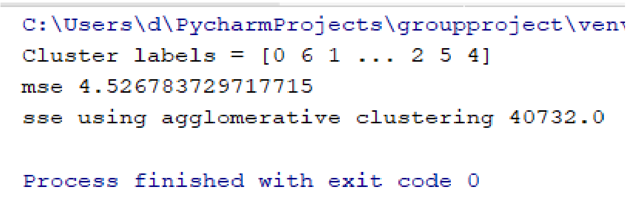


Fig 5.1 The various clusters in the Individual dataset



```
C:\Users\d\PycharmProjects\groupproject\venv
Cluster labels = [0 6 1 ... 2 5 4]
mse 4.526783729717715
sse using agglomerative clustering 40732.0

Process finished with exit code 0
```

Fig 5.2 The MSE and SSE given by Agglomerative clustering

b) Clustering using DBScan

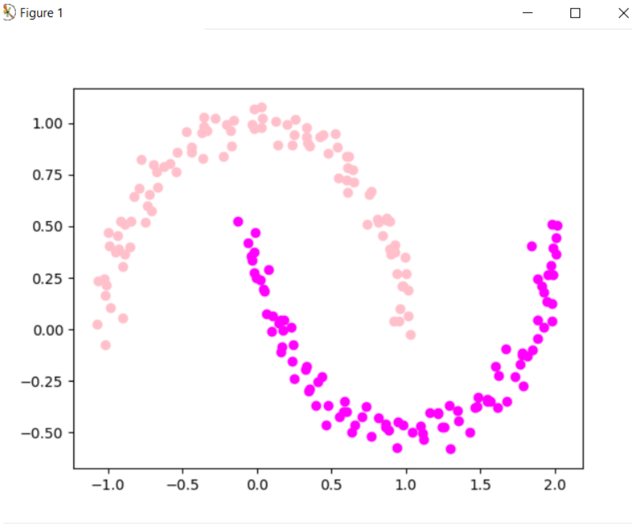
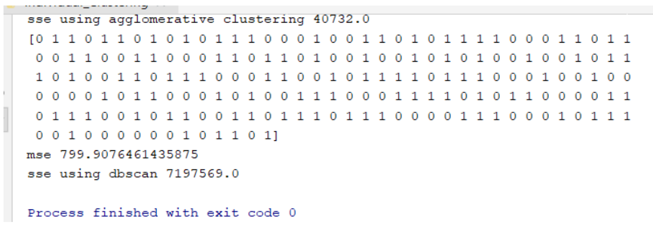


Fig 5.3 The make_moons method used to depict the clustering



```
sse using agglomerative clustering 40732.0
[0 1 1 0 1 1 0 1 0 1 0 1 1 1 0 0 0 1 0 0 1 1 0 1 0 1 1 1 1 0 0 0 1 1 0 1 1
 0 0 1 1 0 0 1 1 0 0 0 1 1 0 1 1 0 1 0 0 1 0 0 1 0 1 0 1 0 0 1 0 0 1 0 1 1
 1 0 1 0 0 1 1 0 1 1 1 0 0 0 1 1 0 0 1 0 1 1 1 1 0 1 1 1 0 0 0 1 0 0 1 0 0
 0 0 0 0 1 0 1 1 0 0 0 1 0 1 0 0 1 1 1 0 0 0 1 1 1 1 0 1 0 1 1 0 0 0 0 1 1
 0 1 1 1 0 0 1 0 1 1 0 0 1 1 0 1 1 1 0 1 1 1 0 0 0 0 1 1 1 0 0 0 1 0 1 1 1
 0 0 1 0 0 0 0 0 0 1 0 1 1 0 1]
mse 799.9076461435875
sse using dbscan 7197569.0

Process finished with exit code 0
```

Fig 5.4 shows the MSE and the SSE

c) Clustering using elbow method and the respective graph with the number of clusters and the WCSS is generated.

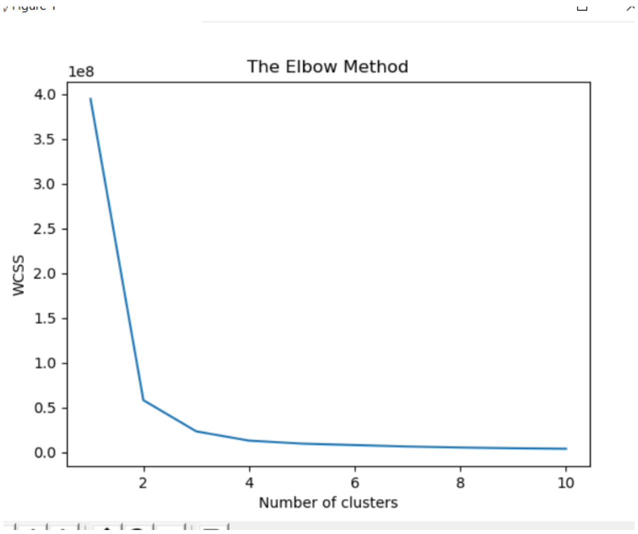
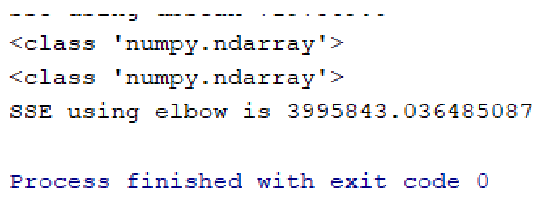


Fig 5.5 Shows the elbow graph

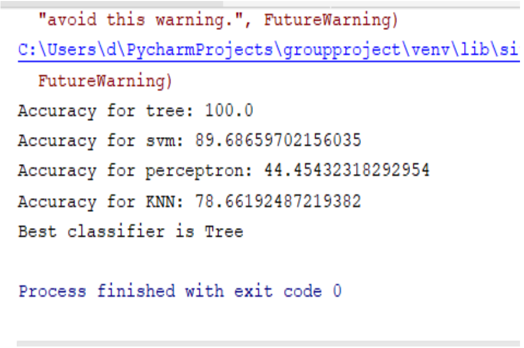


```
<class 'numpy.ndarray'>
<class 'numpy.ndarray'>
SSE using elbow is 3995843.036485087

Process finished with exit code 0
```

Fig 5.6 shows the SSE using the elbow method

6) ANALYSIS FOR DIFFERENT CLASSIFIERS ON INDIVIDUAL DATA:

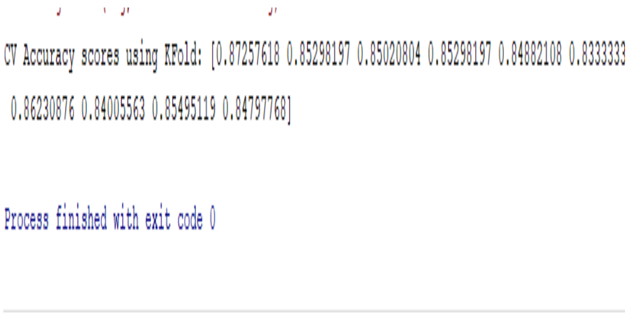


```
  "avoid this warning.", FutureWarning)
C:\Users\d\PycharmProjects\groupproject\venv\lib\si
  FutureWarning)
Accuracy for tree: 100.0
Accuracy for svm: 89.68659702156035
Accuracy for perceptron: 44.45432318292954
Accuracy for KNN: 78.66192487219382
Best classifier is Tree

Process finished with exit code 0
```

Fig 6.1 The best classifier was found to be the tree classifier

7) VALIDATION DONE ON THE INDIVIDUAL DATA:



```
CV Accuracy scores using KFold: [0.87257618 0.85298197 0.85020804 0.85298197 0.84882108 0.8333333
 0.86230876 0.84005563 0.85495119 0.84797768]


Process finished with exit code 0
```

Fig 7.1 The accuracy scores shown

8) CLASSIFICATION:

	Resources Data	**Individual Data**

Agglomerative clustering	Mse=1.3333
	SSE=200
		Mse=4.52
		SSE=40732.0

DBScan	Mse=0.96
	SSE=145
		Mse=799.90
		SSE=7197569.0

Elbow Method	SSE=27.7664	SSE=3995843.0364

Best Classifier	Tree=100
	Svm=20.80
	Perceptron=8.05
	KNN=10.12
		Tree=100
		Svm=89.68
		Perceptron=44.45
		KNN=78.66

Validation accuracies	10 Folds approx. 0.45	10 Folds approx. 0.84

9) RESULTS AND INFERENCE:

Thus we have compared the data for the resources and individual data and based on our analysis we conclude that there is more of carbon emission when individuals make use of the resources and hence due to the huge values and variation we get the results which we have put forth in table 8.1The accuracies of the Individual data seem to be more and better than the other dataset.We have used Clustering,Classifiers,Validation,Regression and used most of the sklearn libraries based on the knowledge gathered in machine learning and have made this comparison chart to see the performance of the various classifiers.