CS F415: DATA MINING
Assignment-3

Team Members :-

1. Ayush Pandey [2015B3A70517H]

2. Abhinav kumar [2015B5A70674H]

3. Shivam Bhagat [2015B5A70460H]

4. Rohan Jain [2015B4A70676H]

Data-set

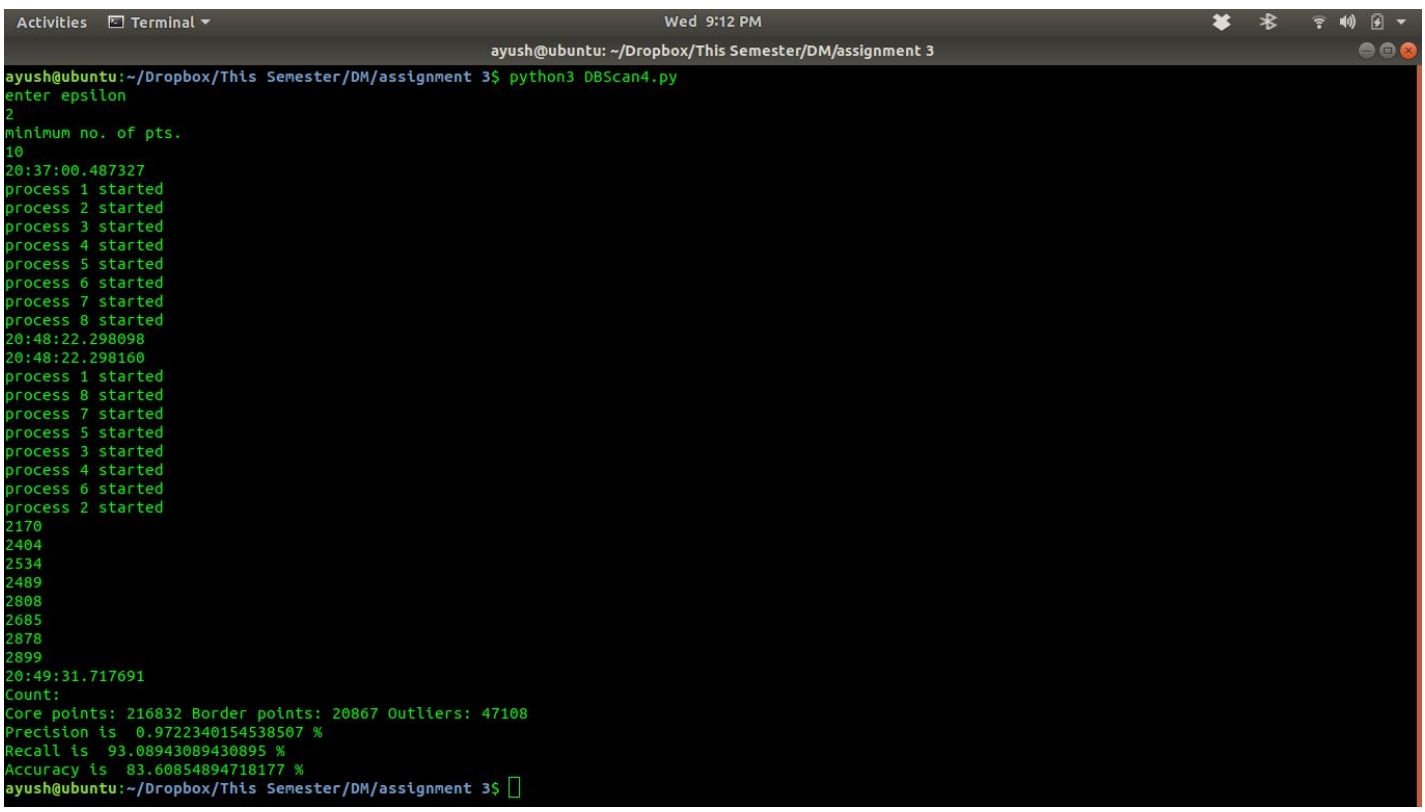
The data-set used is Credit Card Fraud Detection data-set (link) mentioned in assignment description.

DB Scan (Density based Scanning)

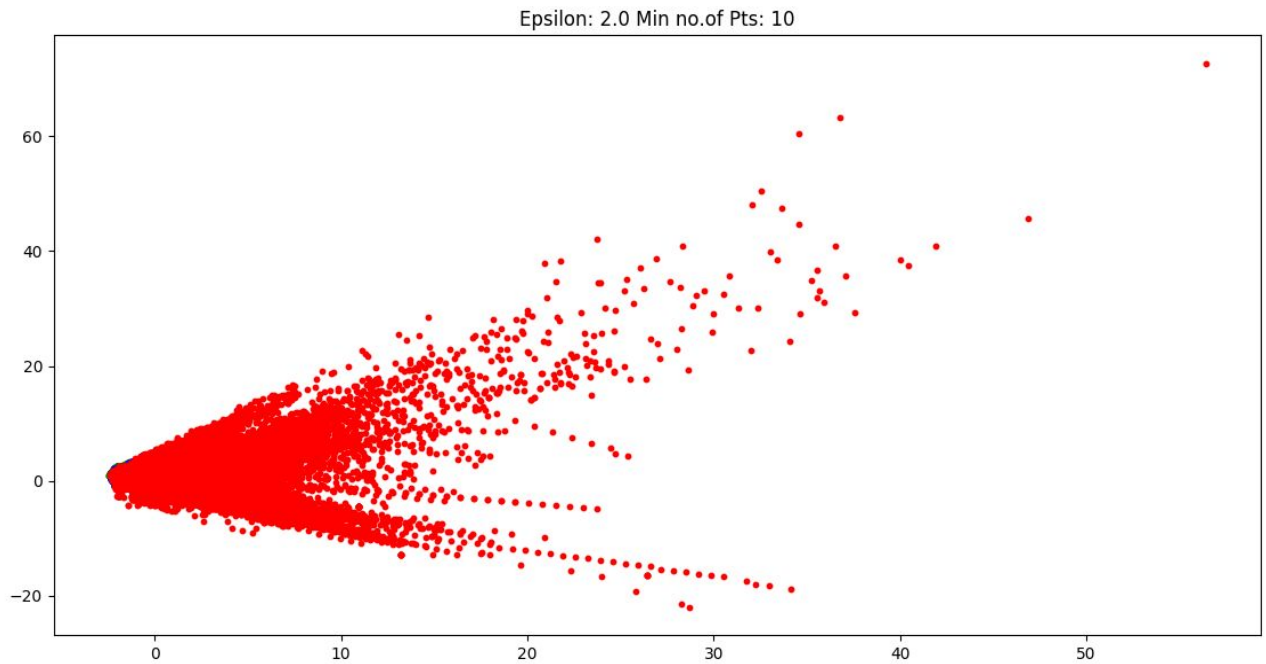
The DB Scan algorithm was implemented for outlier detection with varying values of hyperparameters. Plots of data-set (transformed to 2D) containing predicted core points in green, border points in blue and outlier points in red along with Accuracy and Recall score of 10 runs with varying values of hyperparameters are shown below-

Run 1:

Recall : 83.6% & Accuracy : 93.1.%



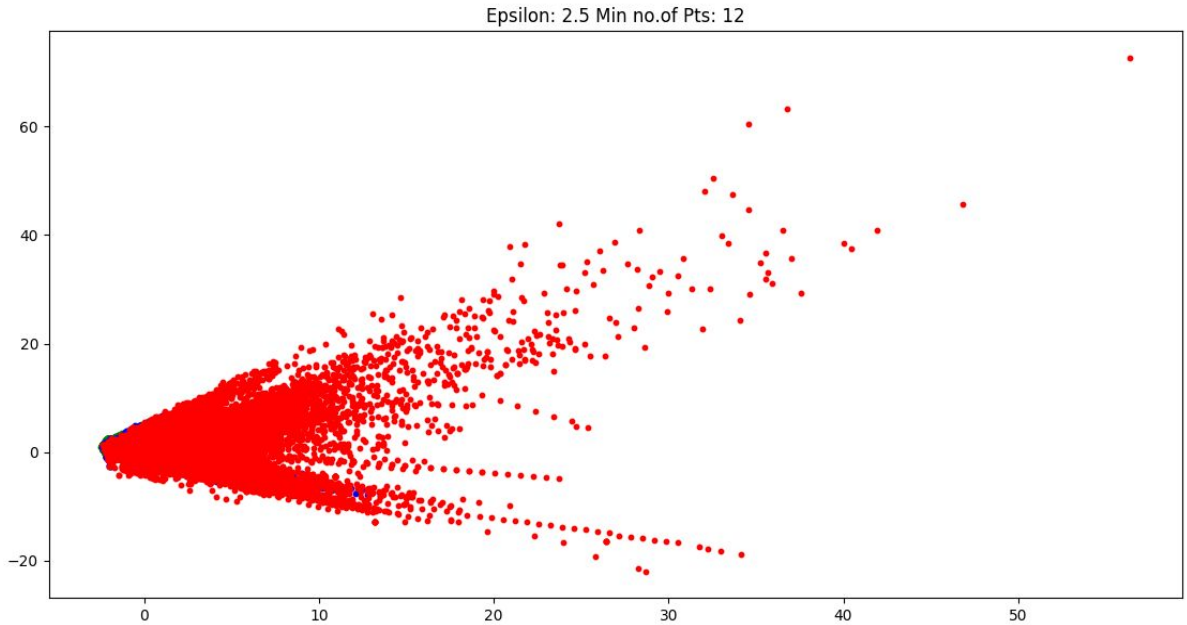
```
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
2
minimum no. of pts.
10
20:37:00.487327
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
20:48:22.298098
20:48:22.298160
process 1 started
process 8 started
process 7 started
process 5 started
process 3 started
process 4 started
process 6 started
process 2 started
2170
2404
2534
2489
2808
2685
2878
2899
20:49:31.717691
Count:
Core points: 216832 Border points: 20867 Outliers: 47108
Precision is  0.9722340154538507 %
Recall is  93.08943089430895 %
Accuracy is  83.60854894718177 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```



Run 2:

Recall : 88.8% & Accuracy : 90.7.%

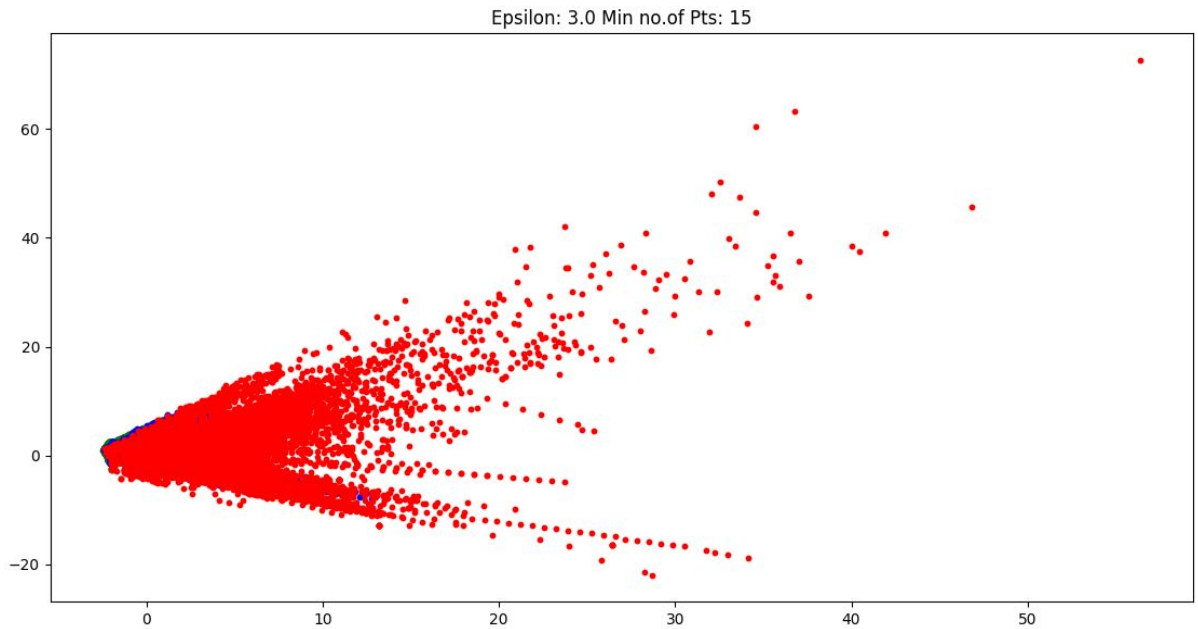
```
Activities   Terminal                                   Wed 9:32 PM
ayush@ubuntu: ~/Dropbox/This Semester/DM/assignment 3
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
2.5
minimum no. of pts.
12
21:12:26.761102
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
21:30:28.065567
21:30:28.065629
process 7 started
process 4 started
process 1 started
process 6 started
process 3 started
process 5 started
process 2 started
process 8 started
1572
1657
1743
1873
2180
2093
2141
2419
21:31:16.261902
Count:
Core points: 242381 Border points: 15678 Outliers: 26748
Precision is  1.6337670106176163 %
Recall is  88.8211382113821 %
Accuracy is  90.7425028176976 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```



Run 3:

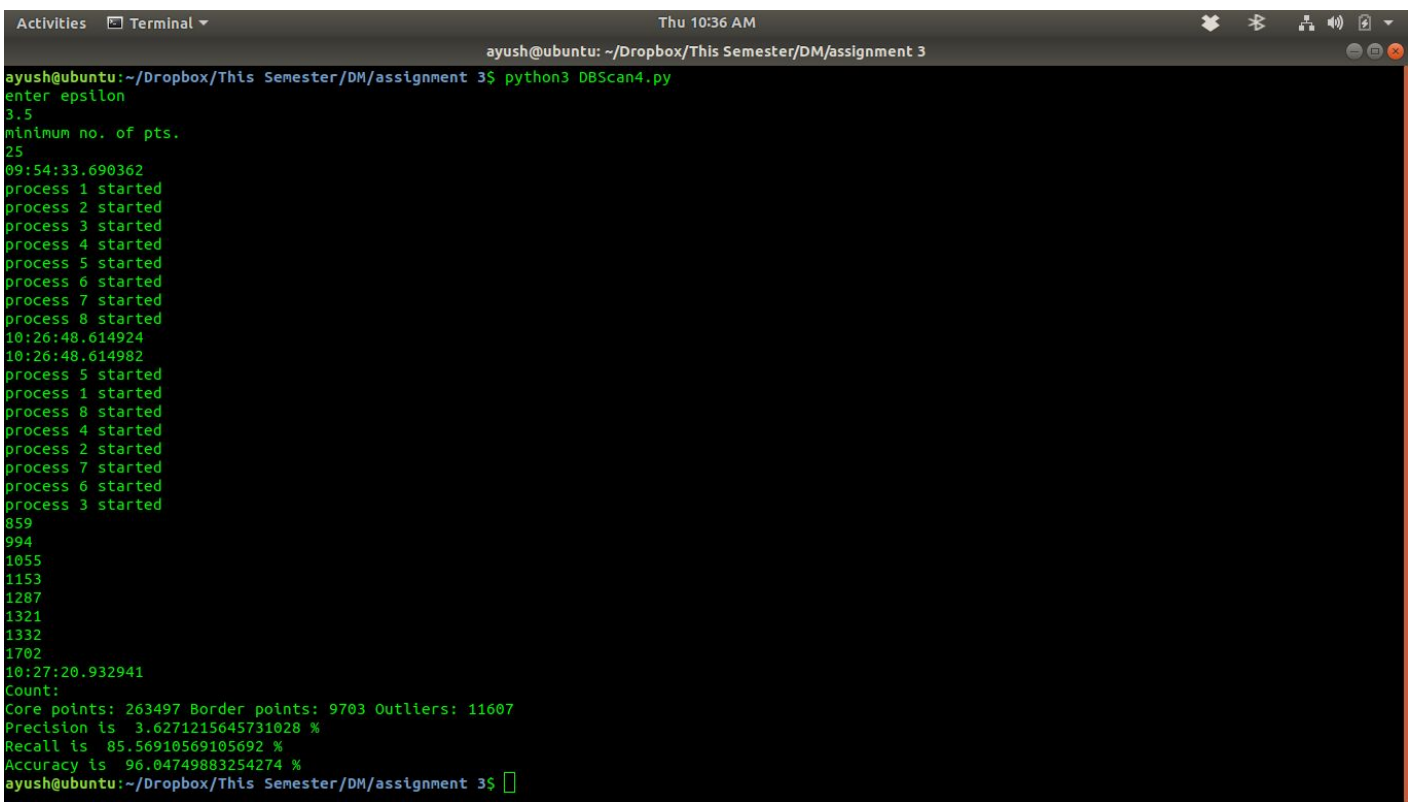
Recall : 86.2% & Accuracy : 94.3.%

```
Activities  Terminal                    Wed 11:16 PM
ayush@ubuntu: ~/Dropbox/This Semester/DM/assignment 3
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
3
minimum no. of pts.
15
21:32:50.155079
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
21:57:46.678789
21:57:46.678865
process 3 started
process 4 started
process 6 started
process 8 started
process 7 started
process 2 started
process 5 started
process 1 started
1096
1181
1197
1297
1473
1522
1552
1912
21:58:23.949231
Count:
Core points: 256880 Border points: 11230 Outliers: 16697
Precision is  2.5393783314367853 %
Recall is  86.1788617886179 %
Accuracy is  94.26243034756871 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```

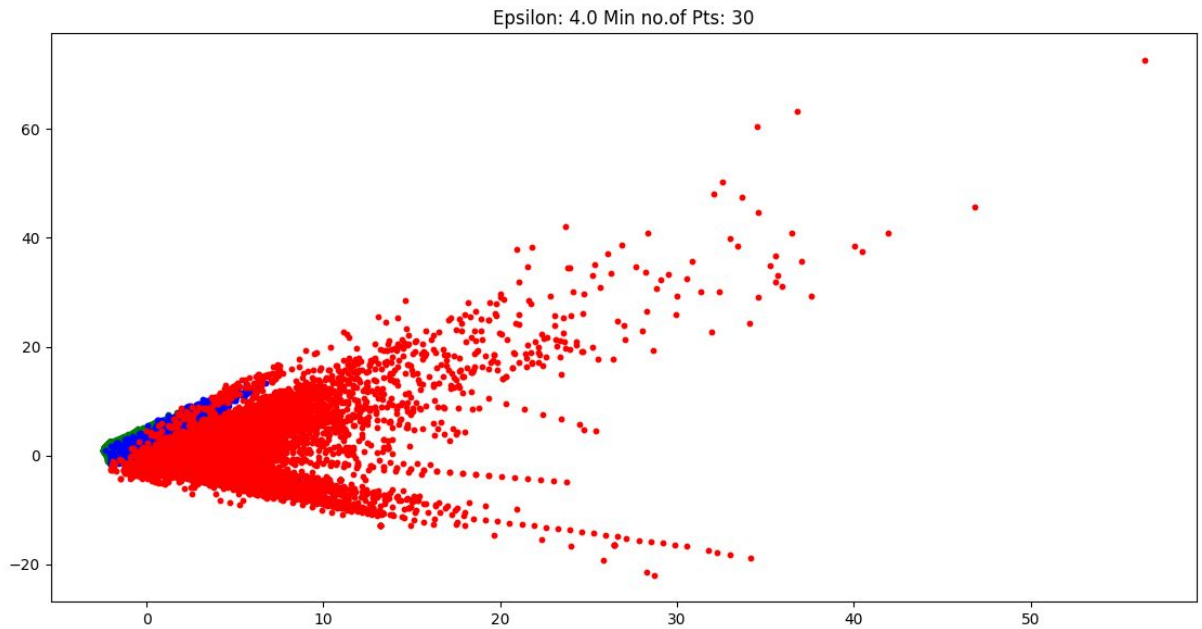


Run 4:

Recall : 85.6% & Accuracy : 96.0.%



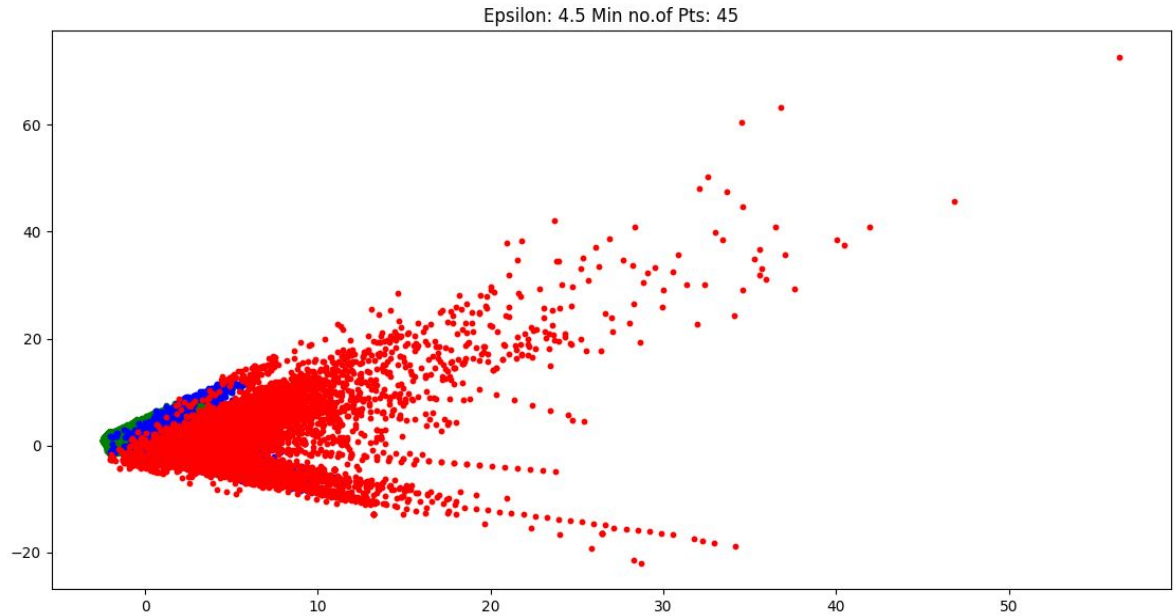
```
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
3.5
minimum no. of pts.
25
09:54:33.690362
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
10:26:48.614924
10:26:48.614982
process 5 started
process 1 started
process 8 started
process 4 started
process 2 started
process 7 started
process 6 started
process 3 started
859
994
1055
1153
1287
1321
1332
1702
10:27:20.932941
Count:
Core points: 263497 Border points: 9703 Outliers: 11607
Precision is  3.6271215645731028 %
Recall is  85.56910569105692 %
Accuracy is  96.04749883254274 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```



Run 5:

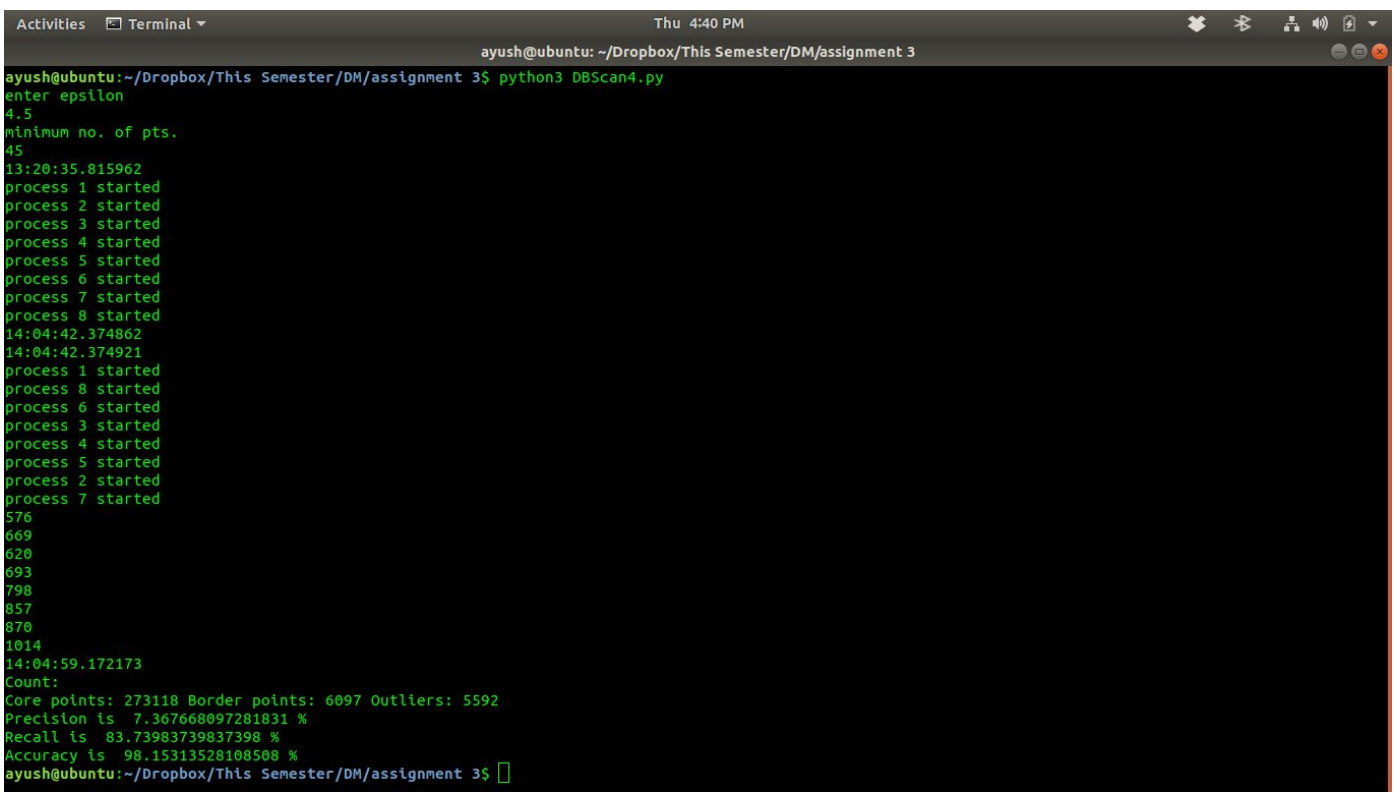
Recall : 84.1% & Accuracy : 97.4.%

```
Activities   Terminal                    Thu 1:20 PM
ayush@ubuntu: ~/Dropbox/This Semester/DM/assignment 3
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
4
minimum no. of pts.
30
10:37:13.943665
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
11:16:10.160149
11:16:10.160209
process 1 started
process 6 started
process 5 started
process 2 started
process 8 started
process 3 started
process 7 started
process 4 started
665
758
743
866
1014
1045
967
1224
11:16:32.686549
Count:
Core points: 269817 Border points: 7282 Outliers: 7708
Precision is  5.371043072132848 %
Recall is  84.14634146341463 %
Accuracy is  97.41158047379453 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```

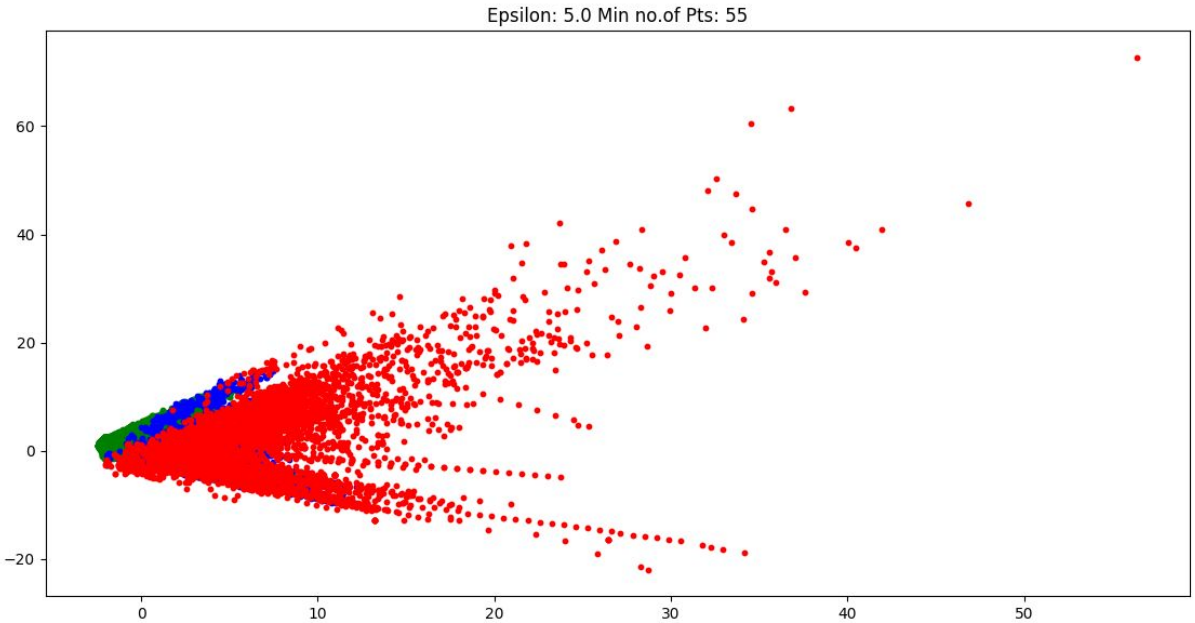


Run 6:

Recall : 83.7% & Accuracy : 98.1.%



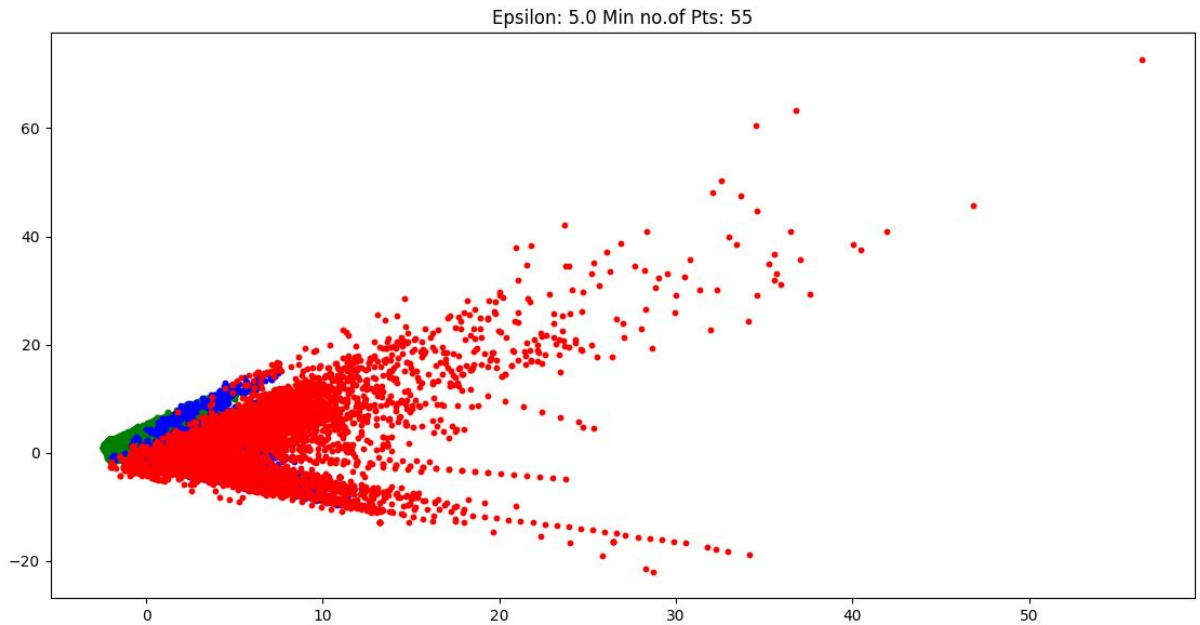
```
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
4.5
minimum no. of pts.
45
13:20:35.815962
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
14:04:42.374862
14:04:42.374921
process 1 started
process 8 started
process 6 started
process 3 started
process 4 started
process 5 started
process 2 started
process 7 started
576
669
620
693
798
857
870
1014
14:04:59.172173
Count:
Core points: 273118 Border points: 6097 Outliers: 5592
Precision is  7.367668097281831 %
Recall is  83.73983739837398 %
Accuracy is  98.15313528108508 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```



Run 7:

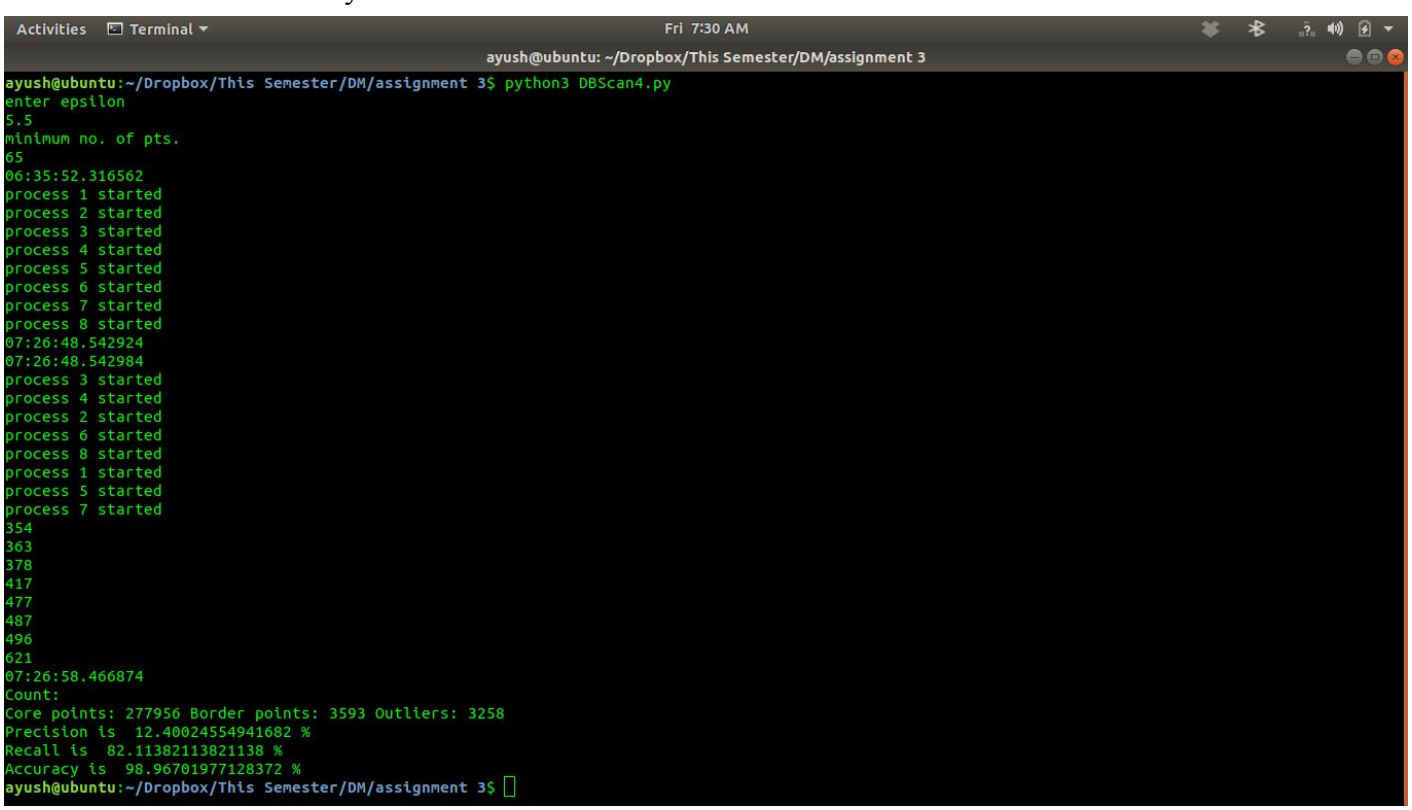
Recall : 83.1% & Accuracy : 98.6.%

```
Activities   Terminal                    Fri 6:35 AM
ayush@ubuntu: ~/Dropbox/This Semester/DM/assignment 3
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
5
minimum no. of pts.
55
05:40:48.679016
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
06:29:07.973290
06:29:07.973354
process 7 started
process 6 started
process 3 started
process 8 started
process 4 started
process 1 started
process 5 started
process 2 started
446
480
489
524
609
677
621
794
06:29:20.899802
Count:
Core points: 275953 Border points: 4640 Outliers: 4214
Precision is  9.705742762221167 %
Recall is  83.130081300813 %
Accuracy is  98.63486501385148 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```

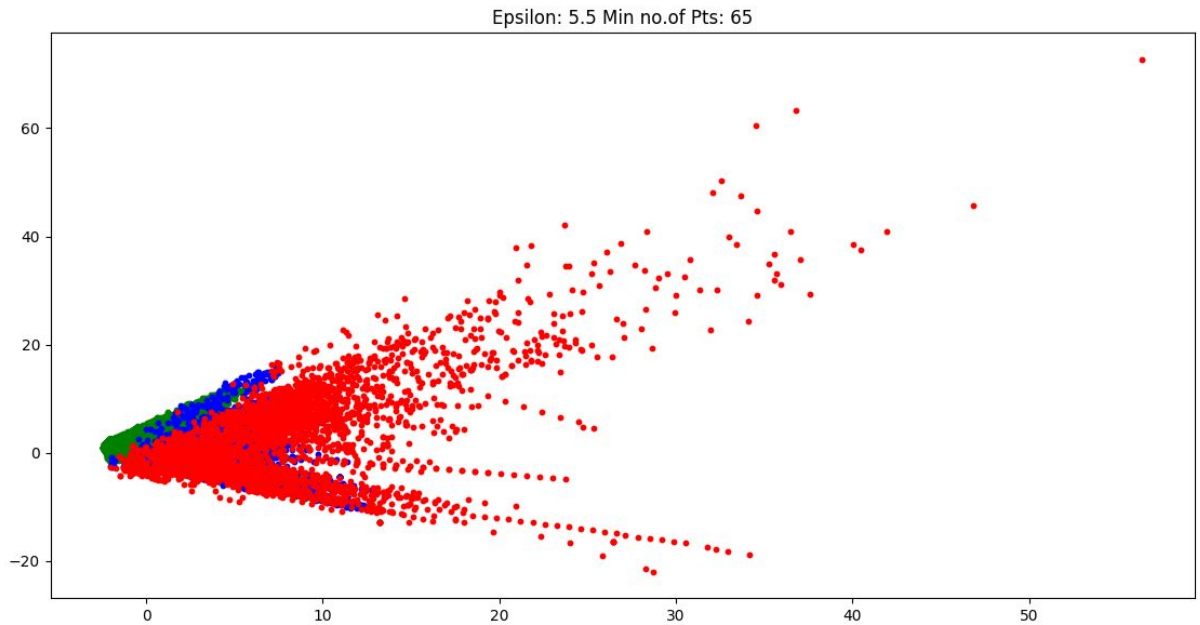


Run 8:

Recall : 82.1% & Accuracy : 98.9.%



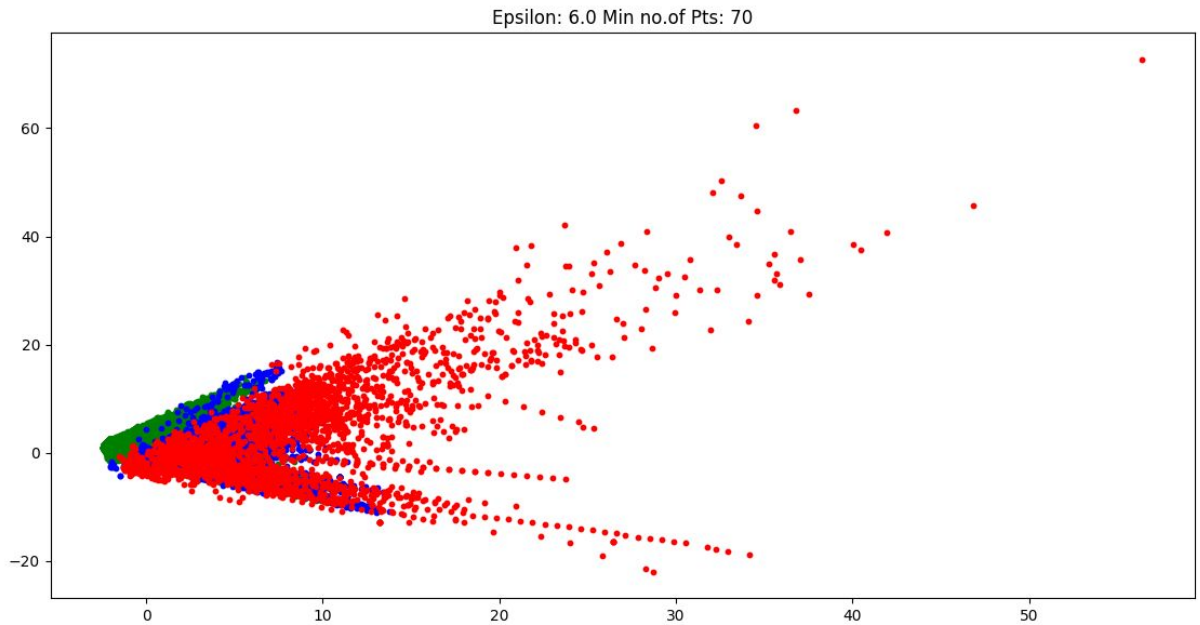
```
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
5.5
minimum no. of pts.
65
06:35:52.316562
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
07:26:48.542924
07:26:48.542984
process 3 started
process 4 started
process 2 started
process 6 started
process 8 started
process 1 started
process 5 started
process 7 started
354
363
378
417
477
487
496
621
07:26:58.466874
Count:
Core points: 277956 Border points: 3593 Outliers: 3258
Precision is  12.40024554941682 %
Recall is  82.11382113821138 %
Accuracy is  98.96701977128372 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```



Run 9:

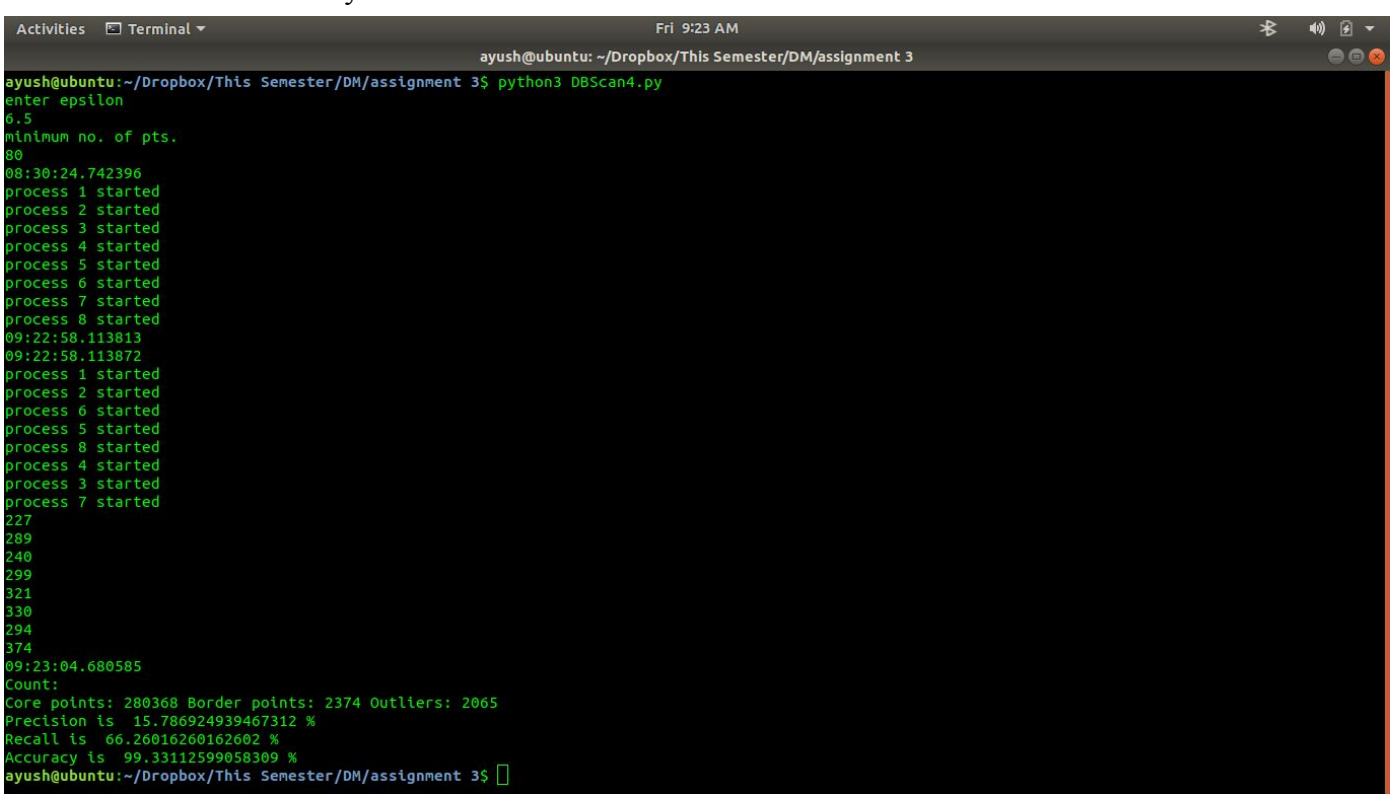
Recall : 81.7% & Accuracy : 99.1.%

```
Activities   Terminal                    Fri 8:30 AM
ayush@ubuntu: ~/Dropbox/This Semester/DM/assignment 3
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
6
minimum no. of pts.
70
07:31:25.206212
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
08:23:33.033016
08:23:33.033081
process 2 started
process 4 started
process 8 started
process 3 started
process 5 started
process 6 started
process 1 started
process 7 started
261
286
342
297
381
358
399
432
08:23:40.547617
Count:
Core points: 279416 Border points: 2756 Outliers: 2635
Precision is  15.2561669829222 %
Recall is  81.70731707317073 %
Accuracy is  99.18435993497351 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
```

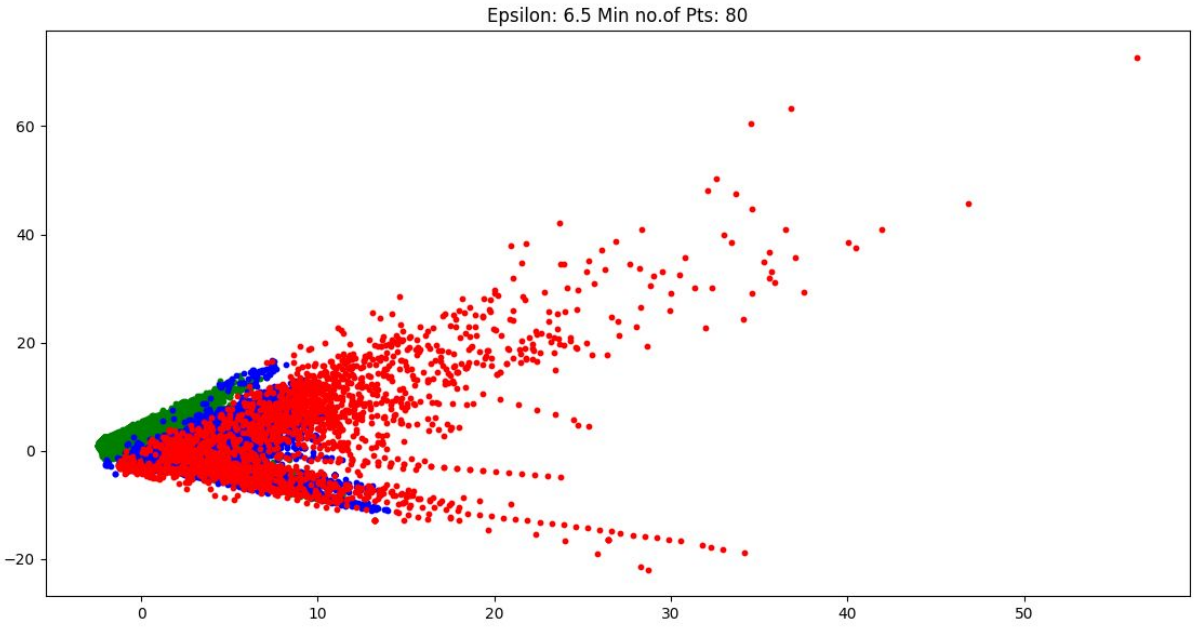


Run 10:

Recall : 66.3% & Accuracy : 99.3.%



```
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$ python3 DBScan4.py
enter epsilon
6.5
minimum no. of pts.
80
08:30:24.742396
process 1 started
process 2 started
process 3 started
process 4 started
process 5 started
process 6 started
process 7 started
process 8 started
09:22:58.113813
09:22:58.113872
process 1 started
process 2 started
process 6 started
process 5 started
process 8 started
process 4 started
process 3 started
process 7 started
227
289
240
299
321
330
294
374
09:23:04.680585
Count:
Core points: 280368 Border points: 2374 Outliers: 2065
Precision is  15.786924939467312 %
Recall is  66.26016260162602 %
Accuracy is  99.33112599058309 %
ayush@ubuntu:~/Dropbox/This Semester/DM/assignment 3$
```


Epsilon: 6.5 Min no.of Pts: 80
60
40
20
0
−20
0
10
20
30
40
50

LOCAL OUTLIER FACTOR

Hyperparameters:

1. **K-value** **:** The k-th closest distance point from a given point which will be used to set the threshold on the number of points in neighbourhood of a point.
2. **Alpha** : The value of **lof** at which we will declare a point a to be outlier or inlier.

Results:

1. **Number of Points:** 50,0000

runtime	k-value	alpha	precision	recall	Confusion matrix	
4m 37s	20	1	0.28%	81%	120	28
					42161	7691
4m 37s	40	1	0.29%	84%	125	23
					42460	7392
4m 59s	60	1	0.348%	98%	146	2
					42436	7416
5m 5s	80	1	0.347%	100%	148	0
					42418	7434
4m 58s	120	1	0.346%	100%	148	0
					42624	7228

2. **Number of Points:** 1,00,0000

runtime	k-value	alpha	precision	recall	Confusion matrix	
18m	20	1	0.217%	81.6%	182	41
					83502	16275
18m	120	1	0.259%	89.55%	222	1
					85397	14460

3. **Number of Points:** 10,0000

runtime	k-value	alpha	precision	recall	Confusion matrix	
10.65s	10	1	0.4%	89.4%	34	4
					8357	1605
10.65s	10	10	0%	0%	0	36
					48	9914
10.65s	10	5	0%	0%	0	38
					58	9904
10.65s	10	2.5	0%	0%	0	38
					131	9831
10.65s	10	1.25	0.3%	21%	8	30
					1553	7409
10.65	10	1.125	0.39%	47%	18	20
					4523	5439
10.65	10	1.001	0.39%	86%	33	5
					8310	1652