Name : Pawan Suresh Bathe Roll No.:CS18M519

CS6745: Mining Massive DataSets

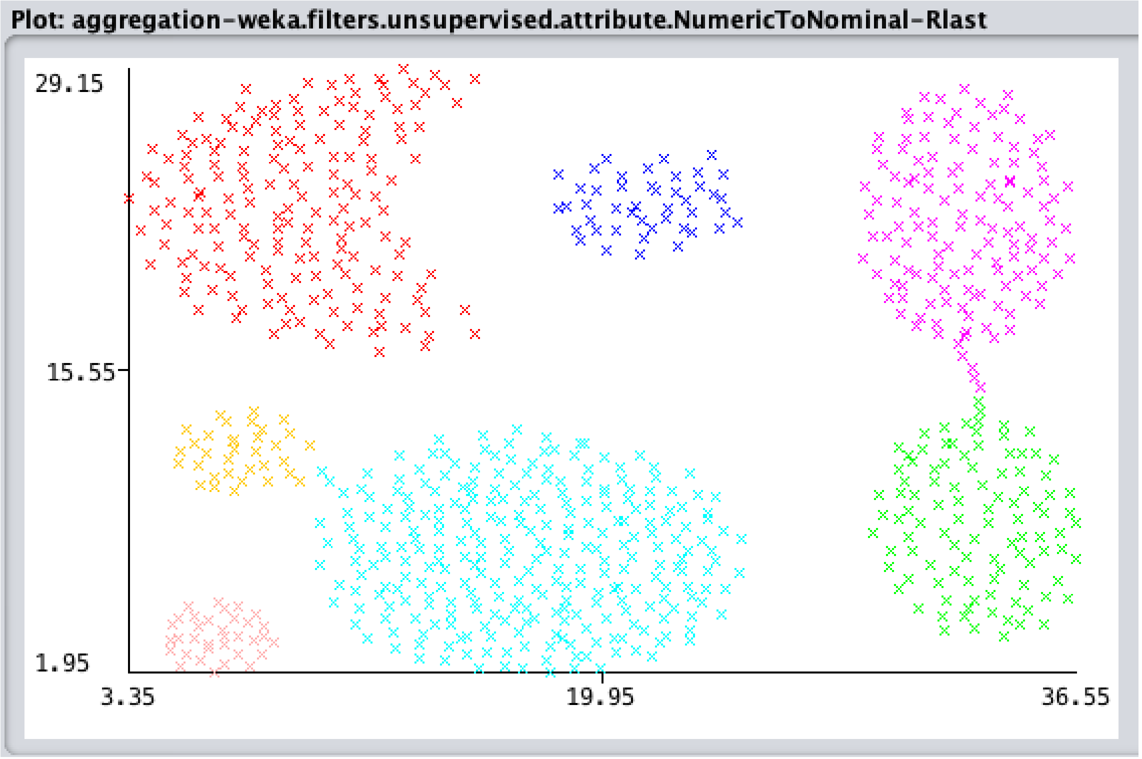
Assignment 3

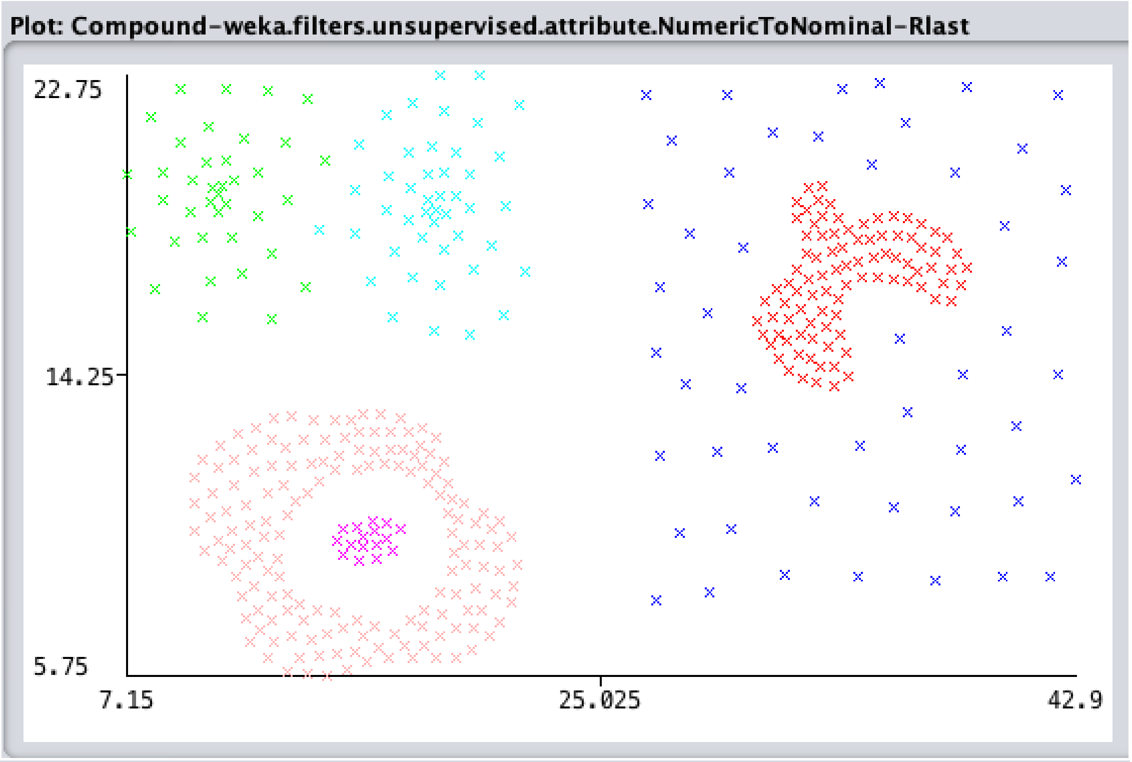
November 22, 2019

1. Convert all 8 datasets into ARFF format.

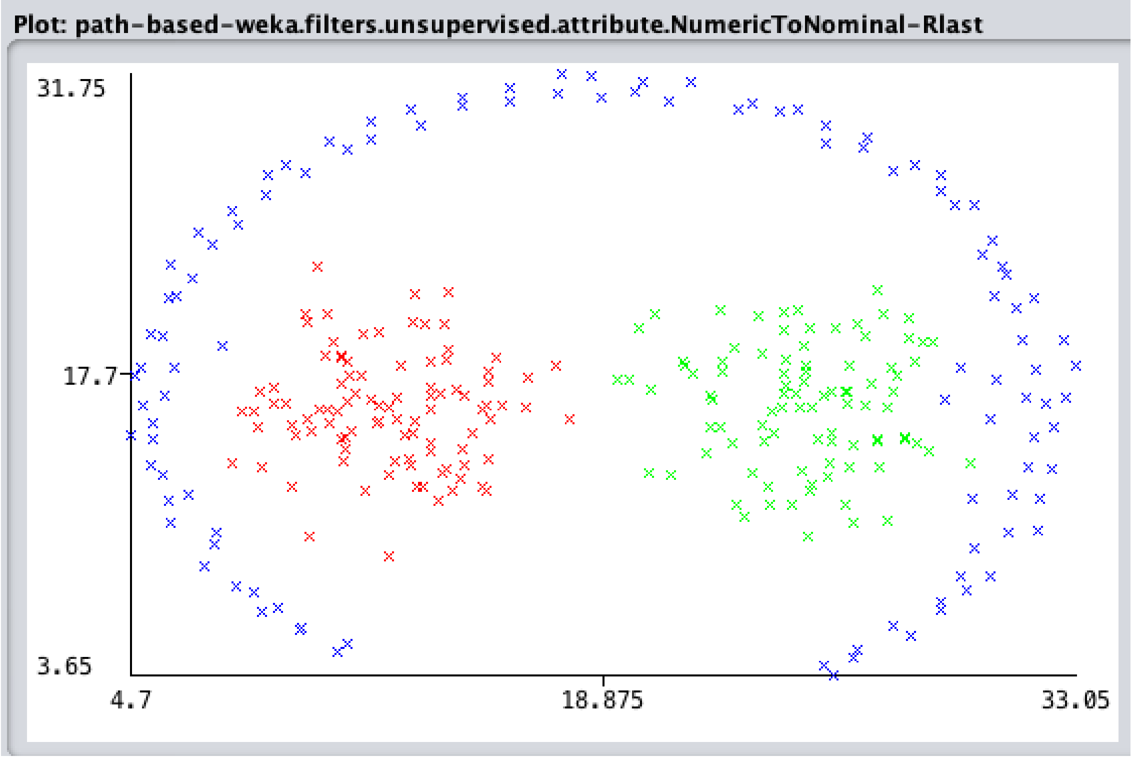
> Answer: All files in ARFF format are stored in Dataset directory.

2. Visualize all 8 datasets. You need to turn in all your plots. Analyze each dataset by visualization and explain how these clustering algorithms will perform in these data (with reasons) : K-means, DBSCAN, hierarchical clustering with single link and complete link.

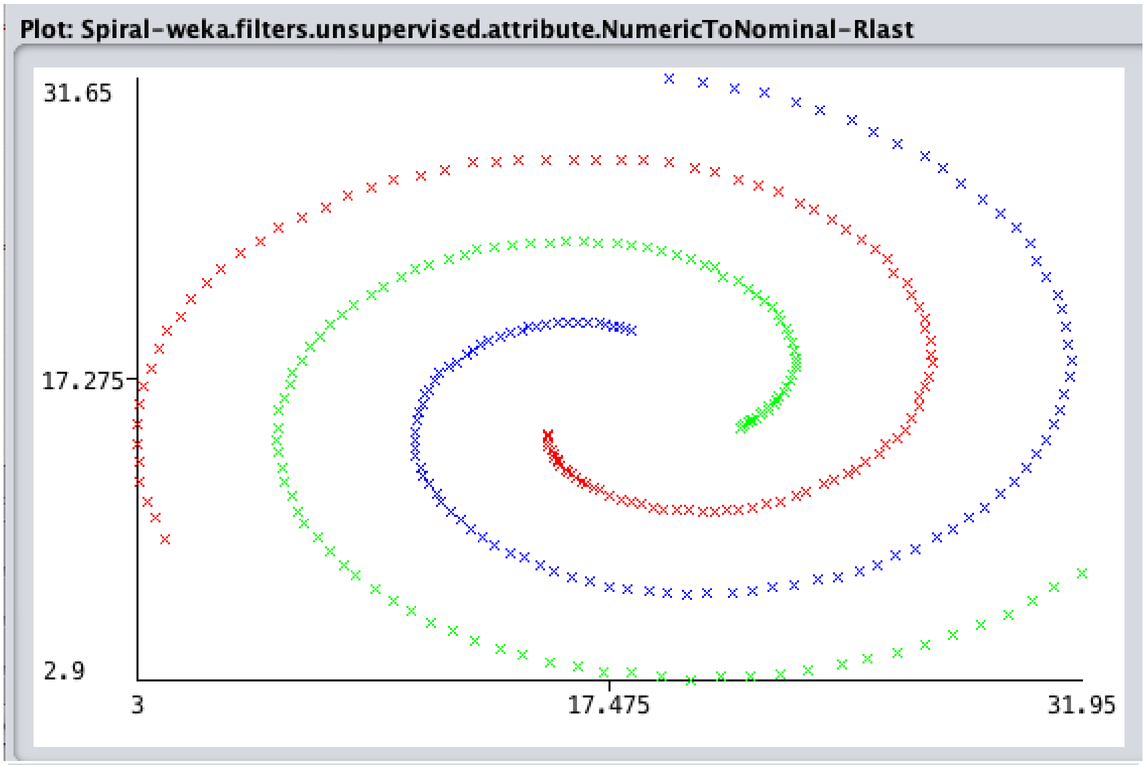




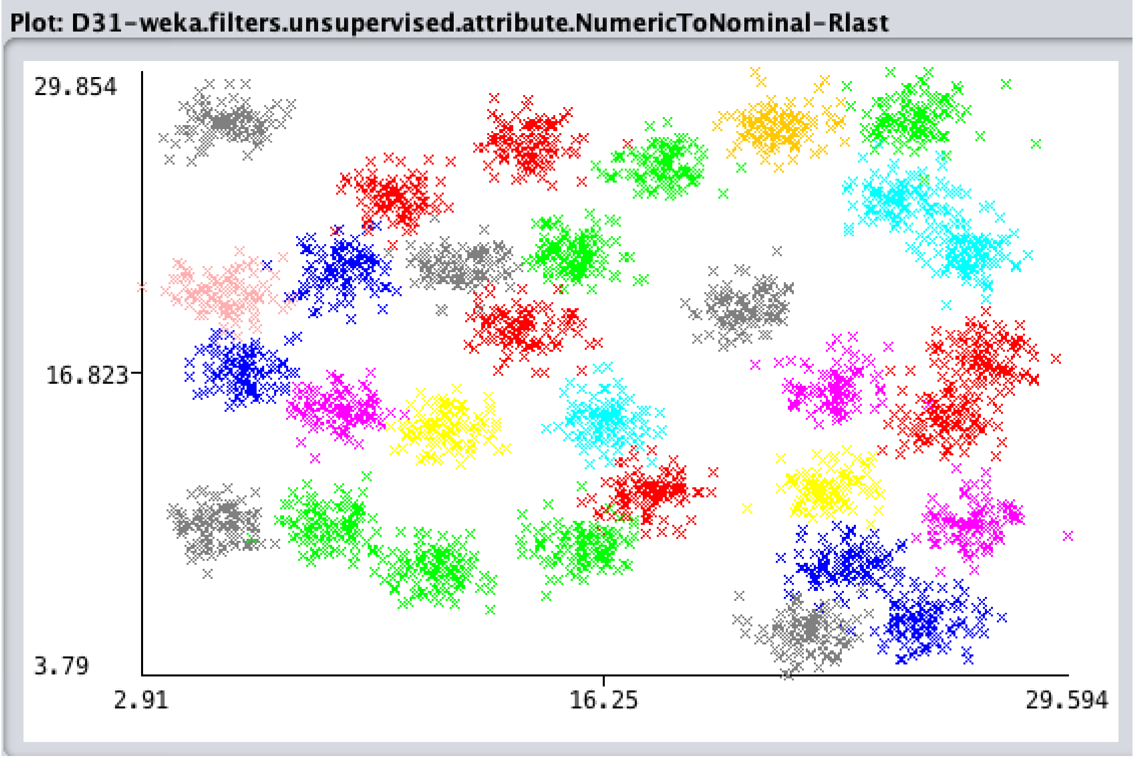
Plot: Compound-weka.filters.unsupervised.attribute.NumericToNominal-Rlast



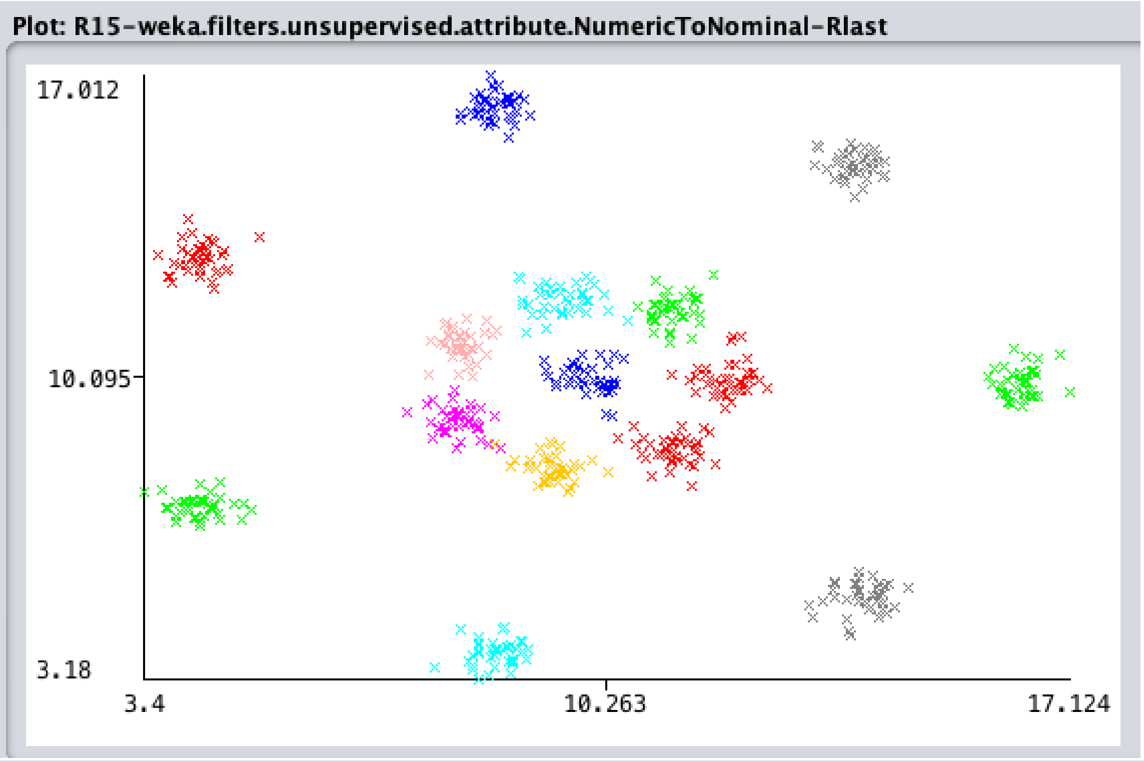
Plot: path-based-weka.filters.unsupervised.attribute.NumericToNominal-Rlast



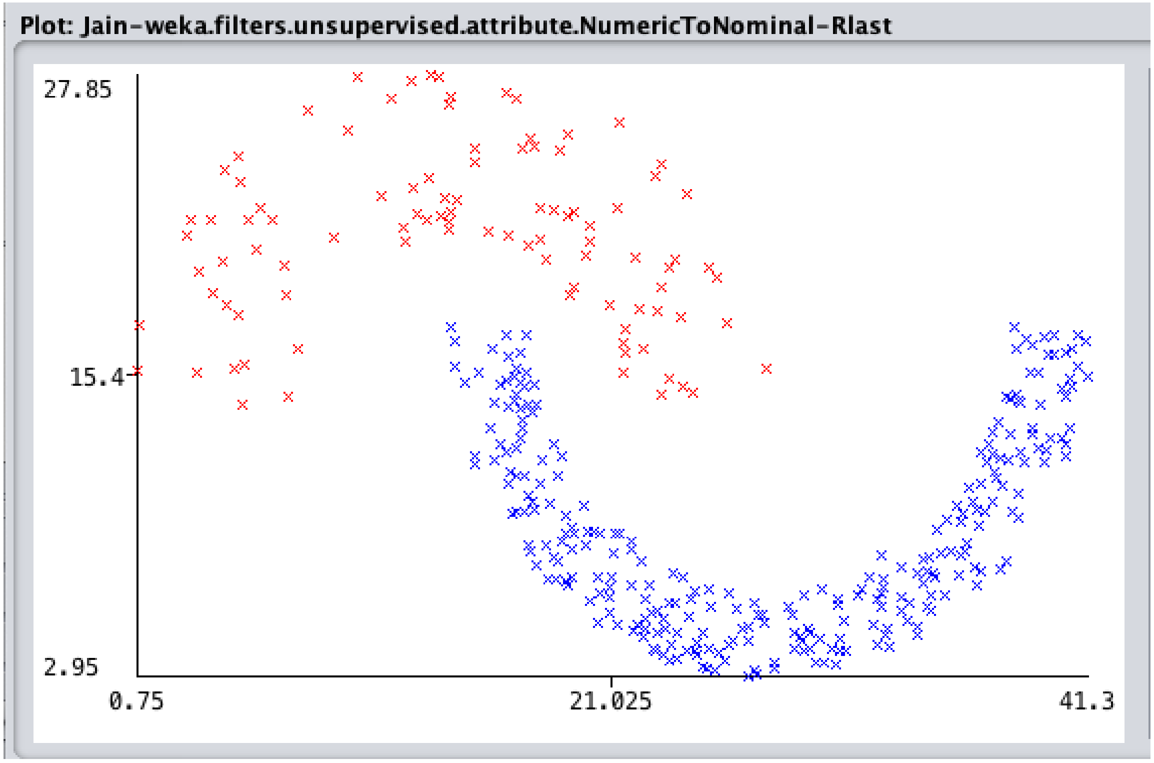
Plot: Spiral-weka.filters.unsupervised.attribute.NumericToNominal-Rlast



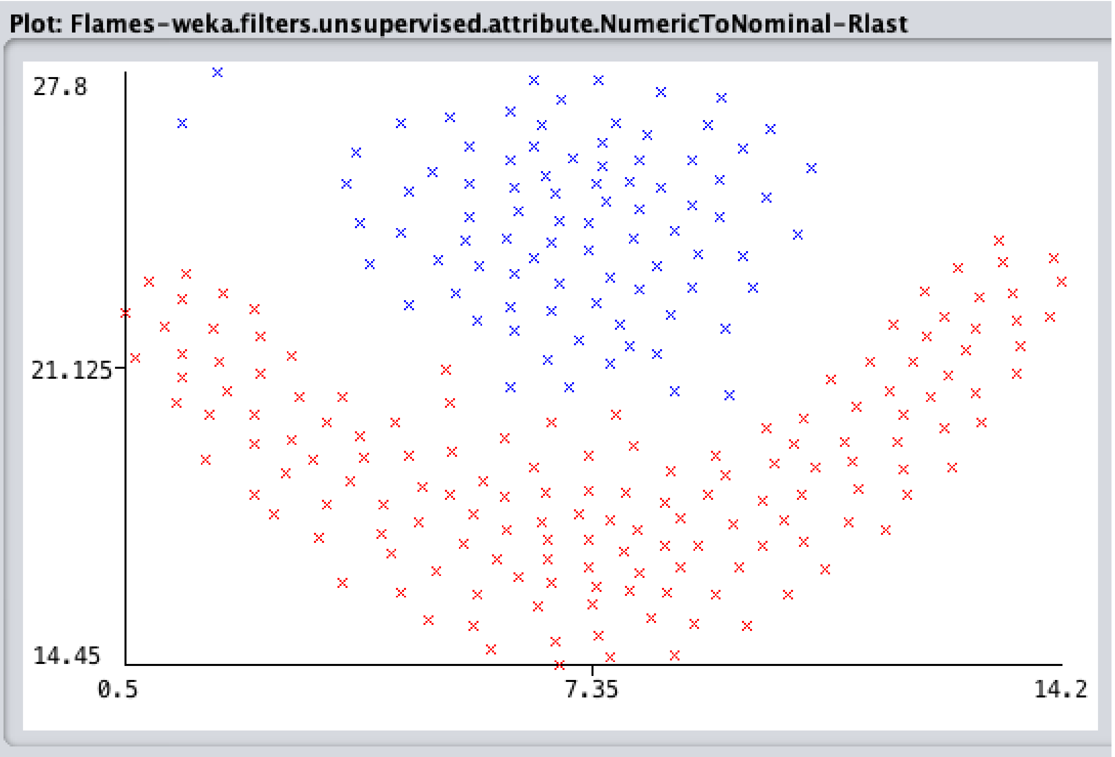
Plot: D31-weka.filters.unsupervised.attribute.NumericToNominal-Rlast



Plot: R15-weka.filters.unsupervised.attribute.NumericToNominal-Rlast



Plot: Jain-weka.filters.unsupervised.attribute.NumericToNominal-Rlast



1. K-means Clustering :
   - Aggregation : K-means doesn't perform very well on this dataset even though clusters are well separated. Since clusters are of variable size bigger clusters tends to get split into smaller clusters.
   - Compound : K-means doesn't work well on compound dataset as well, since there are clusters within clusters and K-means fails to find such non-linear boundaries.
   - Path-based : Even for path-based clustering , K-means doesn't perform well, outer cluster will get split into the inner two cluster on left and right side.
   - Spiral : K-means will not perform well on Spiral dataset as well, since there is bit nesting on spirals and the spirals will get split based on local centroids into multiple clusters.
   - D31: On D31 dataset since cluster points are scatter across different regions, K-means performs poorly since it can't move centroid to completely new region and hence gets stuck in local minima.
   - R15: For the same reason as that of D31 , K-means will not perform well on R15 dataset , it will get stuck in local minima due to scattered clusters.
   - Jain : For Jain data set also the, clusters will get split into multiple smaller clusters , and K-mean will not perform very well.
   - Flames : Same with flames dataset, since clusters are not well separated it doesn't work very well.
2. DBSCAN Clustering:

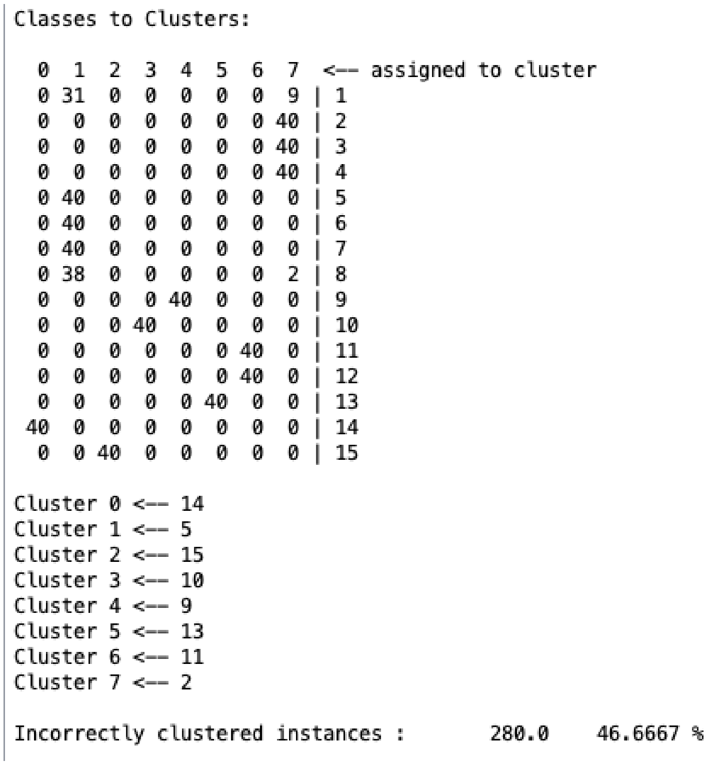
    - Aggregation : DBSCAN will not perform very well on this dataset since some clusters have a points too close to the boundary of another cluster , which might lead DBSCAN to combine those clusters if the point is at epsilon distance from the the point from another clusters boundary
    - Compound : DBSCAN will do OK on this dataset , it will easily able to separate 3 dense clusters , but for other two clusters which are closed and overlapped it will combined them, and last cluster which is surrounding other cluster might largely fall into noise.
    - Path-based: DBSCAN will work very well and will be able to separate all 3 clusters correctly
    - Spiral: Density wise here all spiral clusters are separated, DBSCAN perform excellent on such dataset.
    - D31: DBSCAN will not perform very well on this dataset since the clusters are scattered and shared boundaries with other clusters.
    - R15: DBSCAN won't work well on this dataset as well, it will group centre points into one large clusters and outer points into 7 smaller clusters.
    - Jain: DBSCAN will perform well on this dataset, since density wise they are very well separated.
    - Flames: DBSCAN will perform well on this dataset, since density wise they are very well separated.

  3. Single-Link and Complete Link Hierarchical Clustering:

    - Aggregation: Performs well with complete link Hierarchical Clustering , not so great with single Link clustering, because it stretch another clusters into first cluster
    - Compound: Performs well with both complete link and single Link clustering.
    - Path-based: Performs well with both complete link and single Link clustering.
    - Spiral: Performs well with both complete link and single Link clustering.
    - D31 :
    - R15 : Performs well with complete link Hierarchical Clustering , not so great with single Link clustering, because it stretch another clusters into first cluster
    - Jain : Performs well with complete link Hierarchical Clustering , not so great with single Link clustering, because it stretch another clusters into first cluster
    - Flames : Performs well with complete link Hierarchical Clustering , not so great with single Link clustering, because it stretch another clusters into first cluster

3. Run K-means with R15 dataset. Set k = 8. Report the cluster purity. Vary the value of k from 1 to 20 and study the effect of k on cluster purity. Plot a graph which explains your study.

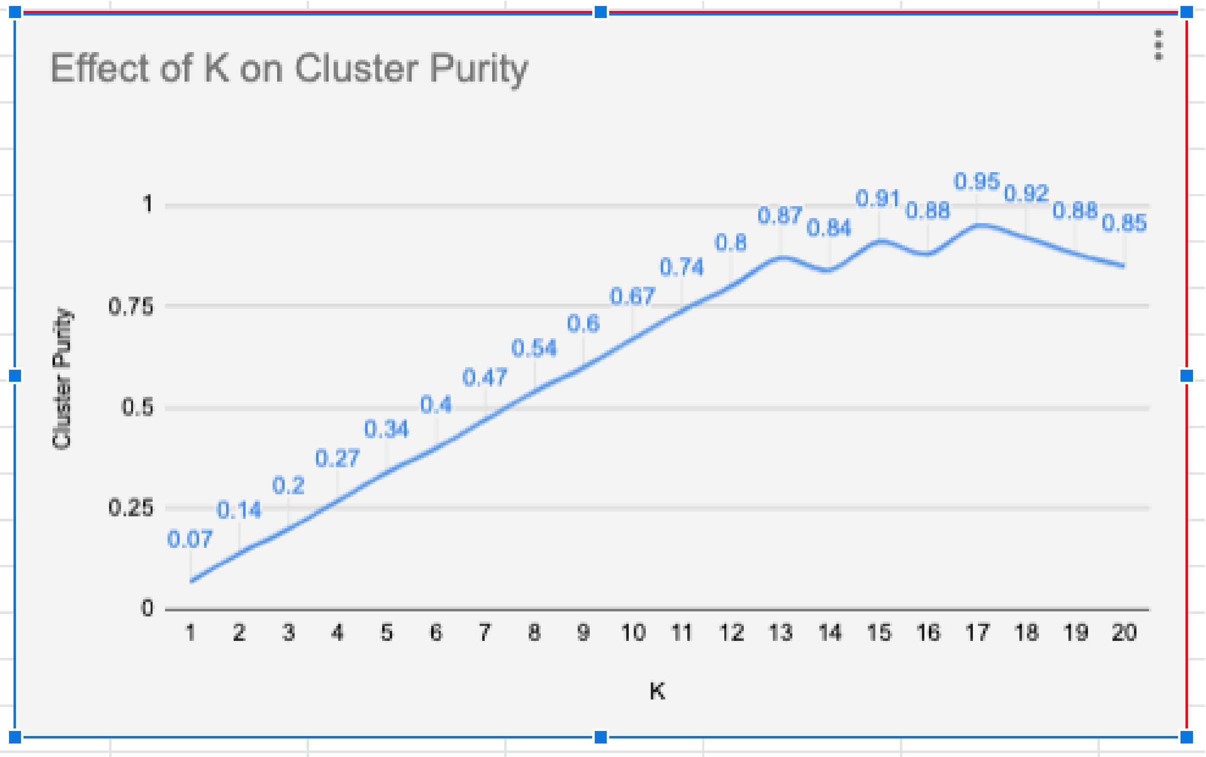
Answer:



```
Classes to Clusters:

  0  1  2  3  4  5  6  7  <-- assigned to cluster
  0 31  0  0  0  0  0  9 | 1
  0  0  0  0  0  0  0 40 | 2
  0  0  0  0  0  0  0 40 | 3
  0  0  0  0  0  0  0 40 | 4
  0 40  0  0  0  0  0  0 | 5
  0 40  0  0  0  0  0  0 | 6
  0 40  0  0  0  0  0  0 | 7
  0 38  0  0  0  0  0  2 | 8
  0  0  0  0 40  0  0  0 | 9
  0  0  0 40  0  0  0  0 | 10
  0  0  0  0  0  0 40  0 | 11
  0  0  0  0  0  0 40  0 | 12
  0  0  0  0  0 40  0  0 | 13
 40  0  0  0  0  0  0  0 | 14
  0  0 40  0  0  0  0  0 | 15

Cluster 0 <-- 14
Cluster 1 <-- 5
Cluster 2 <-- 15
Cluster 3 <-- 10
Cluster 4 <-- 9
Cluster 5 <-- 13
Cluster 6 <-- 11
Cluster 7 <-- 2

Incorrectly clustered instances :       280.0    46.6667 %
```

From confusion matrix, Cluster Purity = (40+40+40+40+40+40+40+40)/600 = 320/600 = 0.53 Also, cluster purity = 1 - incorrectly classified instances = 1 - (46.67/100)= 0.53

K	Incorrectly classified instances %	Cluster Purity
1	93.33	0.07
2	86.67	0.14
3	80	0.20
4	73.33	0.27
5	66.67	0.34
6	60	0.40
7	53.33	0.47
8	46.67	0.54
9	40	0.60
10	33.33	0.67
11	26.67	0.74
12	20	0.80
13	13.83	0.87
14	16	0.84
15	9.33	0.91
16	12.16	0.88
17	5.33	0.95
18	8.5	0.92
19	11.83	0.88
20	14.83	0.85



4. Run DBSCAN with Jain dataset. Again report cluster purity. Study the effect of minpoints and epsilon on cluster purity.

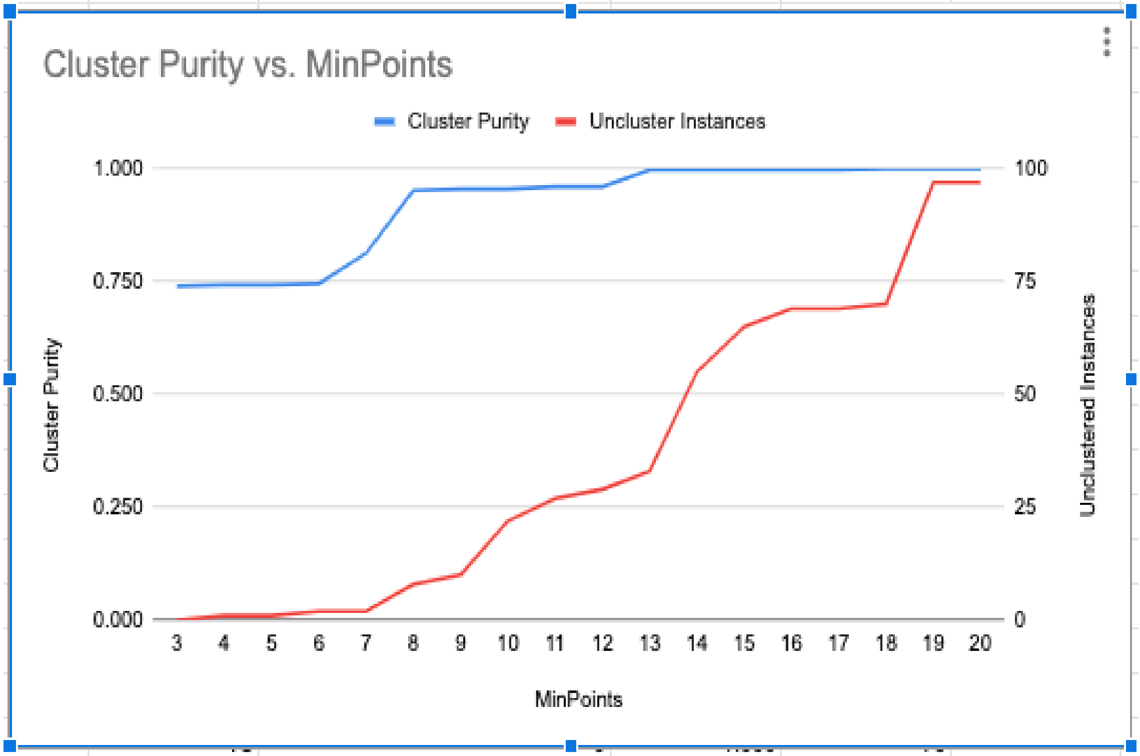
Answer: With default parameter values for minpoints = 6 and Epsilon = 0.9 , Incorrectly clustered instances : 26.0054 % So, cluster purity = 1 - 26.00/100 = 0.74. But this is not very helpful clustering since , it had clustered all data into one cluster and since we have unbalanced dataset (one cluster is bigger than other), the purity is still high.

As we decrease the epsilon , the cluster purity increases, but if we decrease it beyond the certain, the DBSCAN fails to form clusters at all. Here, For Epsilon: 0.09; minPoints: 14 we get best cluster purity = 0.99

Also, for a given Epsilon if we increase min points , cluster purity increases but if we go beyond certain value it fails to form proper clusters. For minPoints = 14 we get best cluster purity.

Below is cluster purity for different value of MinPoints with Epsilon = 0.09

MinPoints	Wrong clustered instances %	Cluster Purity	Unclustered Instances
3	26	0.74	0
4	25.73	0.74	1
5	25.73	0.74	1
6	25.46	0.75	2
7	18.76	0.81	2
8	4.82	0.95	8
9	4.55	0.95	10
10	4.55	0.95	22
11	4.00	0.96	27
12	4.00	0.96	29
13	0.26	0.99	33
14	0.26	0.99	55
15	0.26	0.99	65
16	0.26	0.99	69
17	0.26	0.99	69
18	0	1	70
19	0	1	97
20	0	1	97

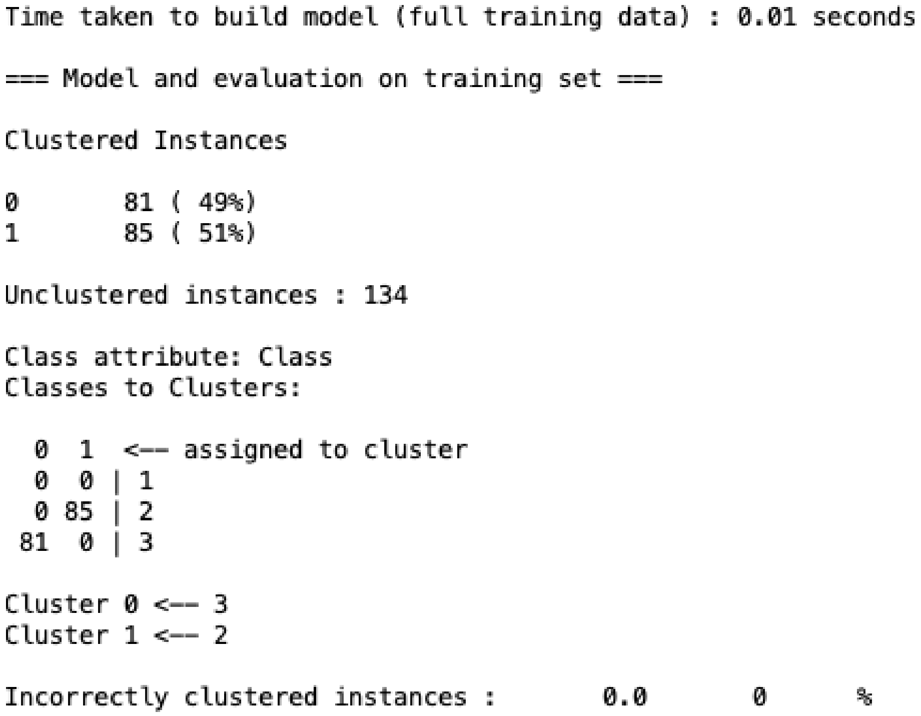


Below is cluster purity for different values of Epsilon with MinPoints=14

Epsilon	Wrong clustered instances %	Cluster Purity	Unclustered Instances
0.07	0	1	83
0.08	0	1	76
0.09	0.26	1	55
0.1	13.94	0.86	31
0.2	26.00	0.74	0
0.3	26.00	0.74	0
0.4	26.00	0.74	0
0.5	26.00	0.74	0
0.6	26.00	0.74	0
0.7	26.00	0.74	0
0.8	26.00	0.74	0
0.9	26.00	0.74	0

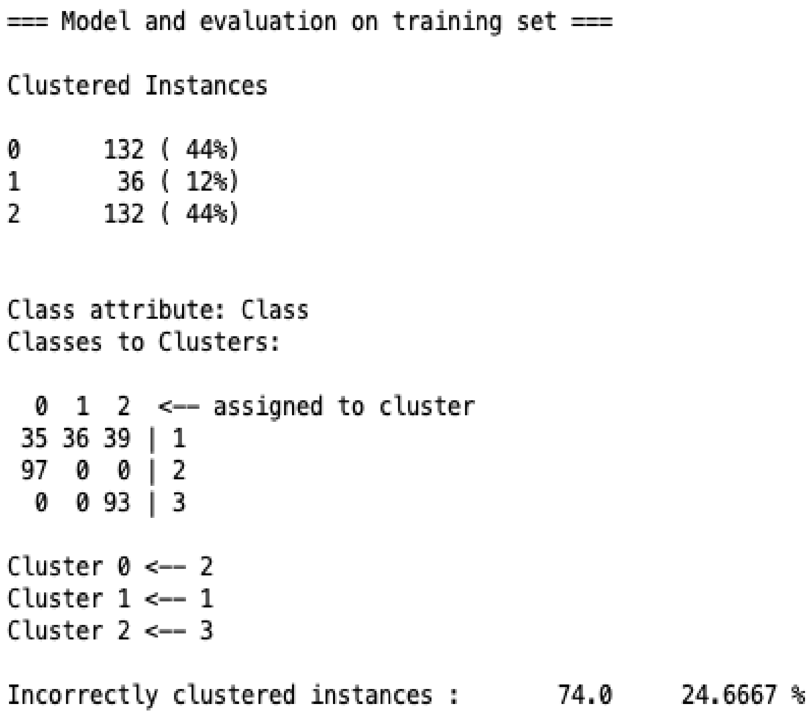
5. Run DBSCAN and hierarchical clustering on Path-based, Spiral and Flames. Compare their performance on each dataset. For hierarchical clustering, you need to experiment with all types of linkages available in Weka to find the one that best suits the data.

Answer: DBSCAN performs well on Path-based dataset: We get cluster purity upto 0.9 .



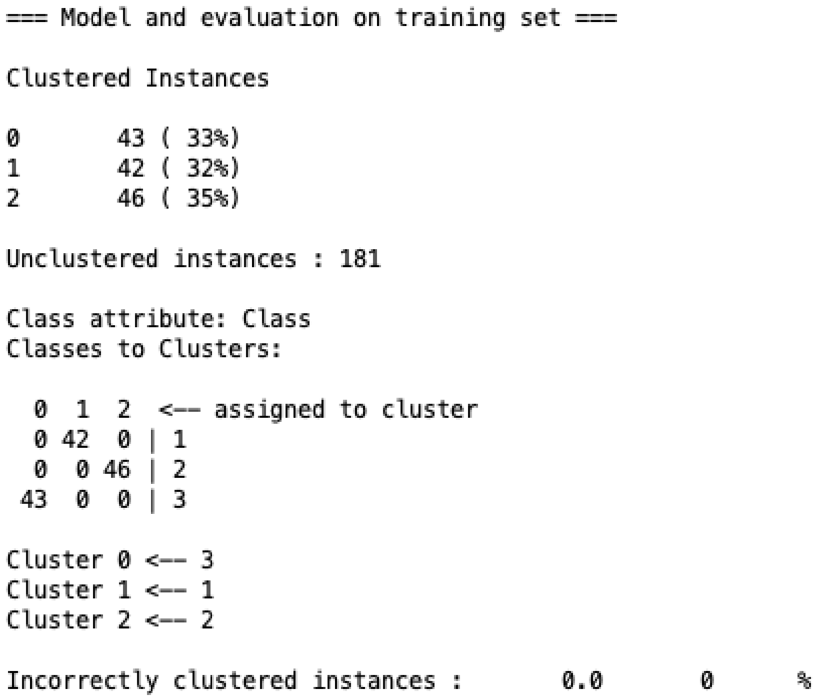
```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       81 ( 49%)
1       85 ( 51%)

Unclustered instances : 134

Class attribute: Class
Classes to Clusters:

  0  1  <-- assigned to cluster
  0  0 | 1
  0 85 | 2
 81  0 | 3

Cluster 0 <-- 3
Cluster 1 <-- 2

Incorrectly clustered instances :       0.0       0      %
```

Hierarchical with ward's linkage works well on Path-based. It is able to cluster data with cluster purity of 0.74



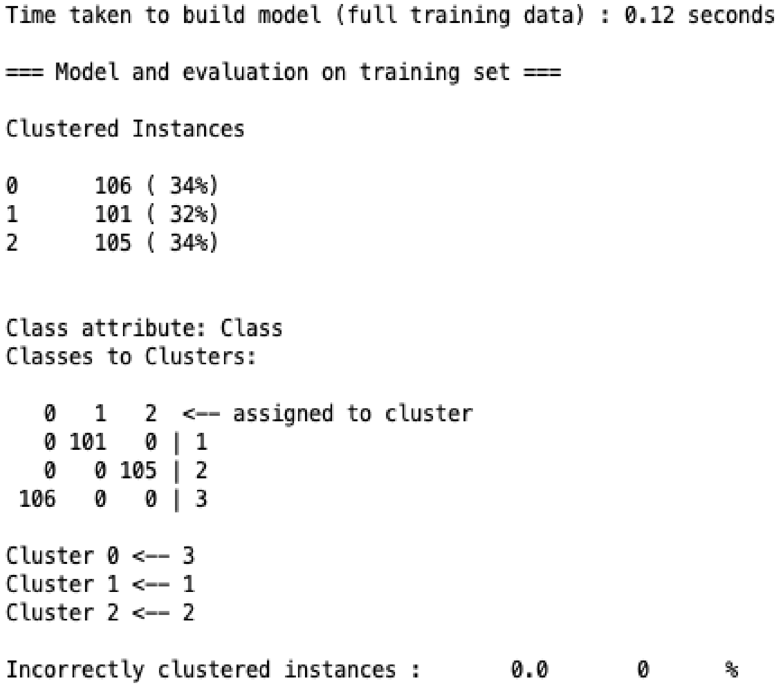
```
=== Model and evaluation on training set ===

Clustered Instances

0      132 ( 44%)
1       36 ( 12%)
2      132 ( 44%)


Class attribute: Class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
 35 36 39 | 1
 97  0  0 | 2
  0  0 93 | 3

Cluster 0 <-- 2
Cluster 1 <-- 1
Cluster 2 <-- 3

Incorrectly clustered instances :       74.0     24.6667 %
```

DBSCAN performs well with Spiral Dataset as well, we get high cluster purity with DBSCAN on Spiral Dataset.



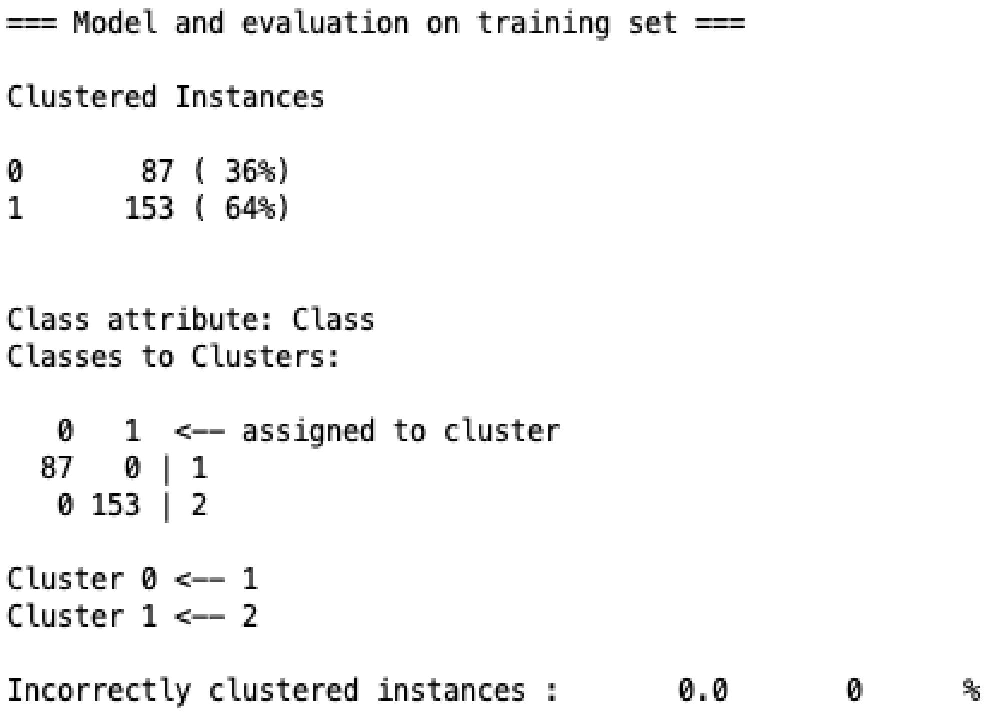
```
=== Model and evaluation on training set ===

Clustered Instances

0       43 ( 33%)
1       42 ( 32%)
2       46 ( 35%)

Unclustered instances : 181

Class attribute: Class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0 42  0 | 1
  0  0 46 | 2
 43  0  0 | 3

Cluster 0 <-- 3
Cluster 1 <-- 1
Cluster 2 <-- 2

Incorrectly clustered instances :       0.0       0      %
```

Hierarchical with single linkage works very on Path-based. It is able to cluster data with cluster purity of 1, with all cluster clearly formed.



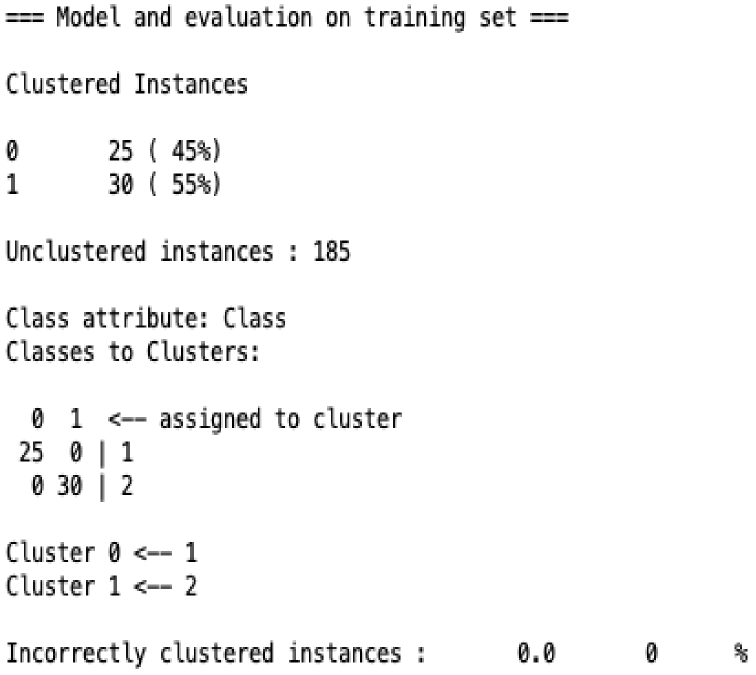
```
Time taken to build model (full training data) : 0.12 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      106 ( 34%)
1      101 ( 32%)
2      105 ( 34%)


Class attribute: Class
Classes to Clusters:

   0   1   2  <-- assigned to cluster
   0 101   0 | 1
   0   0 105 | 2
 106   0   0 | 3

Cluster 0 <-- 3
Cluster 1 <-- 1
Cluster 2 <-- 2

Incorrectly clustered instances :       0.0       0      %
```

DBSCAN performs well with Flames Dataset as well, we get high cluster purity with DBSCAN on Spiral Dataset.



```
=== Model and evaluation on training set ===

Clustered Instances

0       87 ( 36%)
1      153 ( 64%)


Class attribute: Class
Classes to Clusters:

   0   1  <-- assigned to cluster
  87   0 | 1
   0 153 | 2

Cluster 0 <-- 1
Cluster 1 <-- 2

Incorrectly clustered instances :       0.0       0      %
```

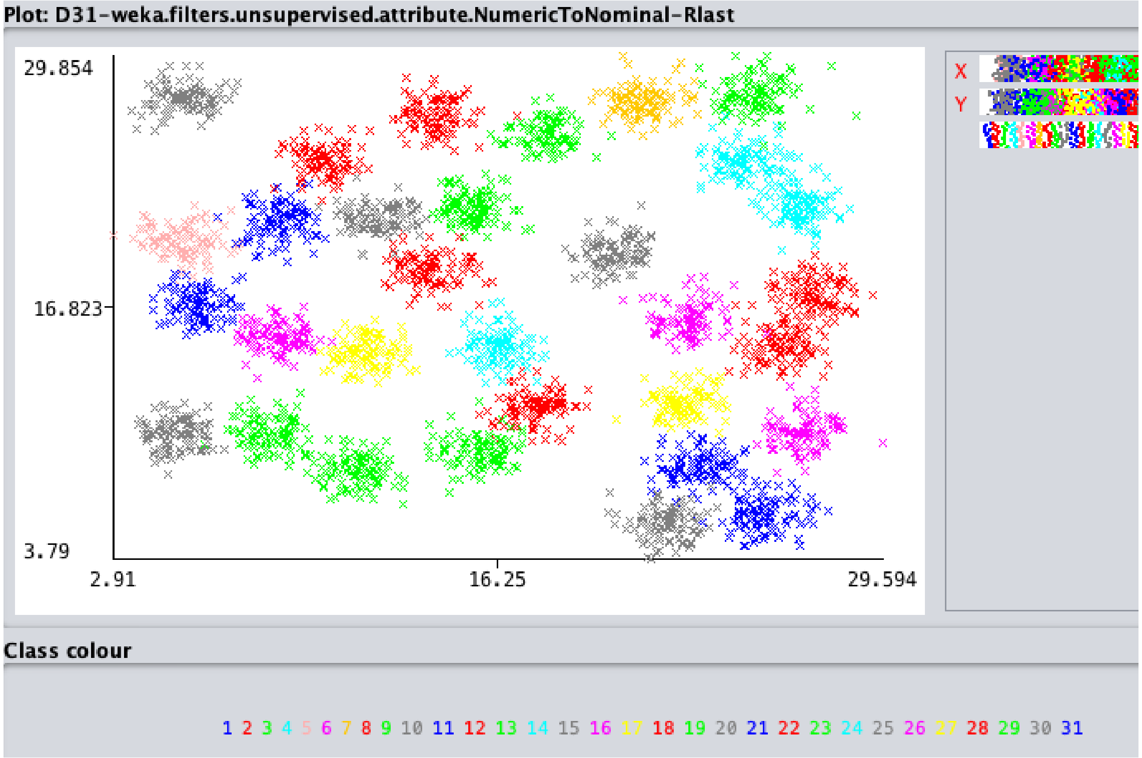
Hierarchical with Wards linkage works very on Path-based. It is able to cluster data with cluster purity of 1, with all cluster clearly formed.

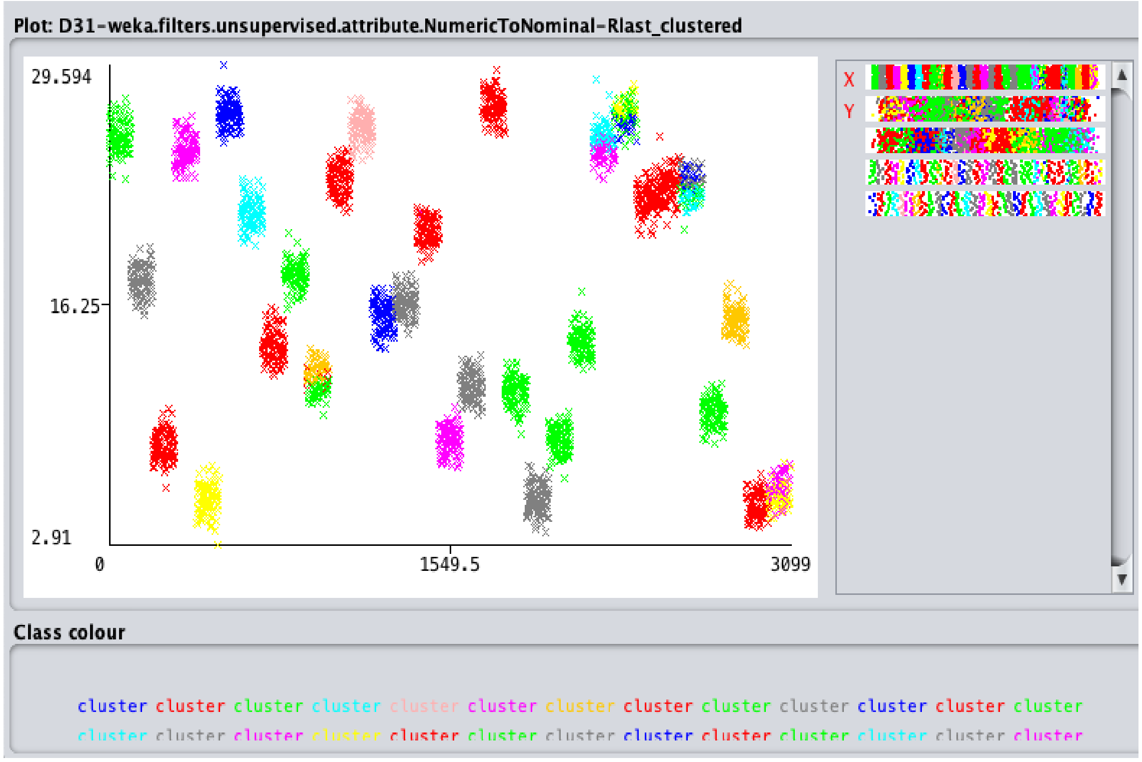


```
=== Model and evaluation on training set ===

Clustered Instances

0      25 ( 45%)
1      30 ( 55%)

Unclustered instances : 185

Class attribute: Class
Classes to Clusters:

  0  1  <-- assigned to cluster
 25  0 | 1
  0 30 | 2

Cluster 0 <-- 1
Cluster 1 <-- 2

Incorrectly clustered instances :      0.0      0     %
```

6. Run K-means with D31 dataset. Can you recover all 31 clusters with k = 32? If not, can you recover all clusters by increasing the value of k? What happens when you apply DBSCAN? Apply hierarchical clustering with Ward's linkage. How does it perform?

Answer: We are able to retrieve 32 clusters with K=32 in D31 but the cluster accuracy is not good. See below original classes and clusters formed using K-means for D31 using K = 32. Increasing value of K is not helping either.




Plot: D31-weka.filters.unsupervised.attribute.NumericToNominal-Rlast_clustered
29.594
16.25
2.91
0
1549.5
3099
X
Y
Class colour
cluster cluster cluster cluster cluster cluster cluster cluster cluster cluster cluster cluster cluster
cluster cluster cluster cluster cluster cluster cluster cluster cluster cluster cluster cluster cluster

D31 dataset performs even worse with DBSCAN with cluster purity ranging nearly less than 0.1 , increasing min points or decreasing epsilon also doesn't help much and ignore lot of points while clustering. D31 performs significantly better with hierarchical clustering with Ward's linkage.