Lab 1 732A75

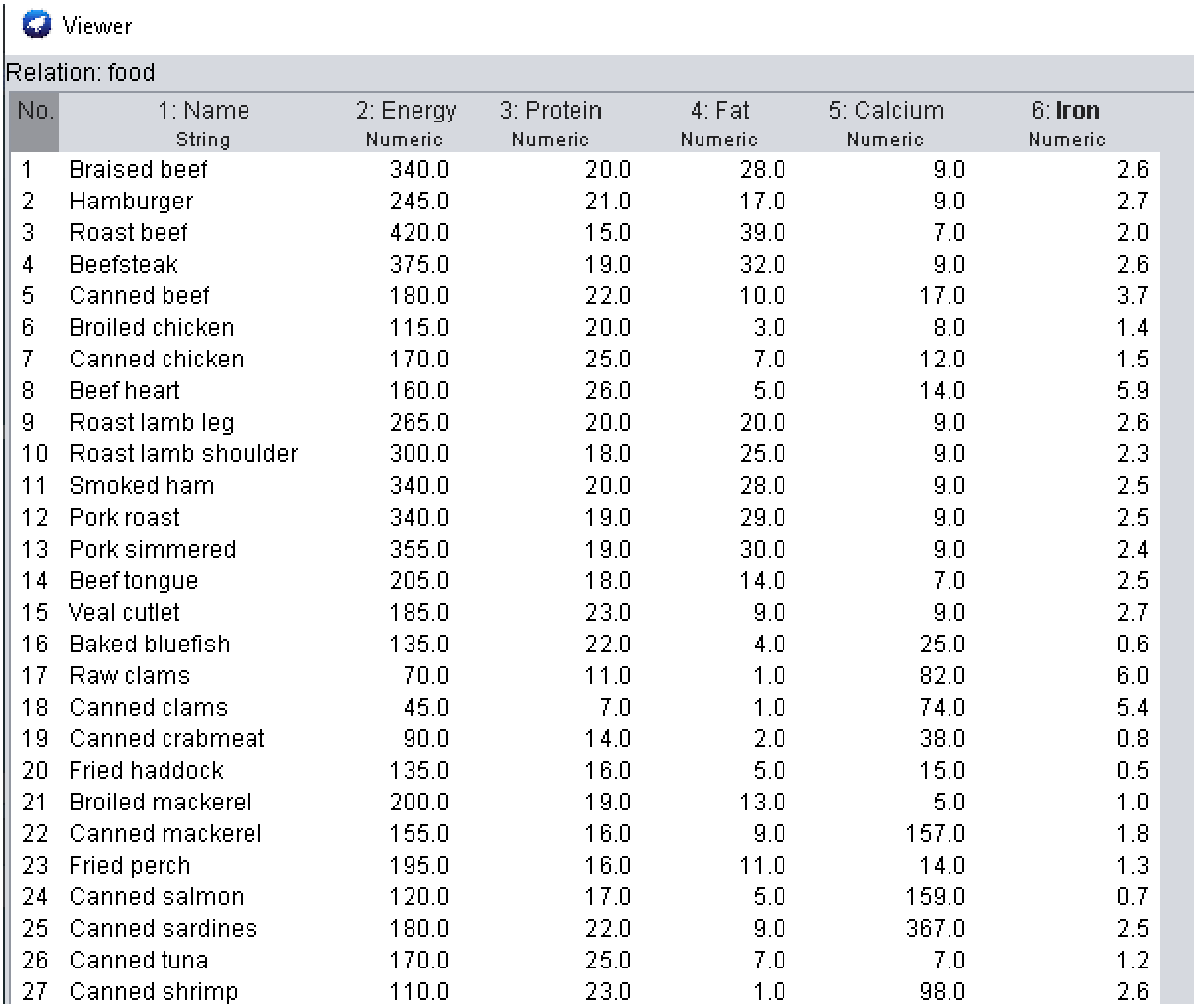
Aman Kumar Nayak (amana551) & Mahmood Siddique (mahsi404)

3/1/2020

About Data

We have 27 measure of Nutrient levels which were measured in a 3 ounce portion of various foods.

Name is the name of the item. **Energy** is the number of calories. **Protein** is the amount of protein in grams. **Fat** is the amount of fat in grams. **Calcium** is the amount of calcium in milligrams. **Iron** is the amount of iron in milligrams.



Relation: food

No.	1: Name String	2: Energy Numeric	3: Protein Numeric	4: Fat Numeric	5: Calcium Numeric	6: **Iron** Numeric
1	Braised beef	340.0	20.0	28.0	9.0	2.6
2	Hamburger	245.0	21.0	17.0	9.0	2.7
3	Roast beef	420.0	15.0	39.0	7.0	2.0
4	Beefsteak	375.0	19.0	32.0	9.0	2.6
5	Canned beef	180.0	22.0	10.0	17.0	3.7
6	Broiled chicken	115.0	20.0	3.0	8.0	1.4
7	Canned chicken	170.0	25.0	7.0	12.0	1.5
8	Beef heart	160.0	26.0	5.0	14.0	5.9
9	Roast lamb leg	265.0	20.0	20.0	9.0	2.6
10	Roast lamb shoulder	300.0	18.0	25.0	9.0	2.3
11	Smoked ham	340.0	20.0	28.0	9.0	2.5
12	Pork roast	340.0	19.0	29.0	9.0	2.5
13	Pork simmered	355.0	19.0	30.0	9.0	2.4
14	Beef tongue	205.0	18.0	14.0	7.0	2.5
15	Veal cutlet	185.0	23.0	9.0	9.0	2.7
16	Baked bluefish	135.0	22.0	4.0	25.0	0.6
17	Raw clams	70.0	11.0	1.0	82.0	6.0
18	Canned clams	45.0	7.0	1.0	74.0	5.4
19	Canned crabmeat	90.0	14.0	2.0	38.0	0.8
20	Fried haddock	135.0	16.0	5.0	15.0	0.5
21	Broiled mackerel	200.0	19.0	13.0	5.0	1.0
22	Canned mackerel	155.0	16.0	9.0	157.0	1.8
23	Fried perch	195.0	16.0	11.0	14.0	1.3
24	Canned salmon	120.0	17.0	5.0	159.0	0.7
25	Canned sardines	180.0	22.0	9.0	367.0	2.5
26	Canned tuna	170.0	25.0	7.0	7.0	1.2
27	Canned shrimp	110.0	23.0	1.0	98.0	2.6

Data Set

Question 1 : SimpleKmeans

1.1: Choose a set of attributes for clustering and give a motivation.

Kmeans algorithm is directly not applicable to the categorical variable as sample space for categorical data is discrete, and thus Euclidean distance function on such space does not provide meaningful results.

Since Name attribute is discrete; thus, we will ignore it while considering remaining continuous numerical values based features; namely Energy, Protein, Fat, Calcium, and Iron.

1.2: Experiment with at least two different numbers of clusters, e.g. 2 and 5, but with the same seed value 10.

Obtained Result at seed value 10

When Number of cluster(K) = 2 with seed 10

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A “weka.core.EuclideanDistance -R first-last” -I 500 -num-slots 1 -S 10

Relation: food

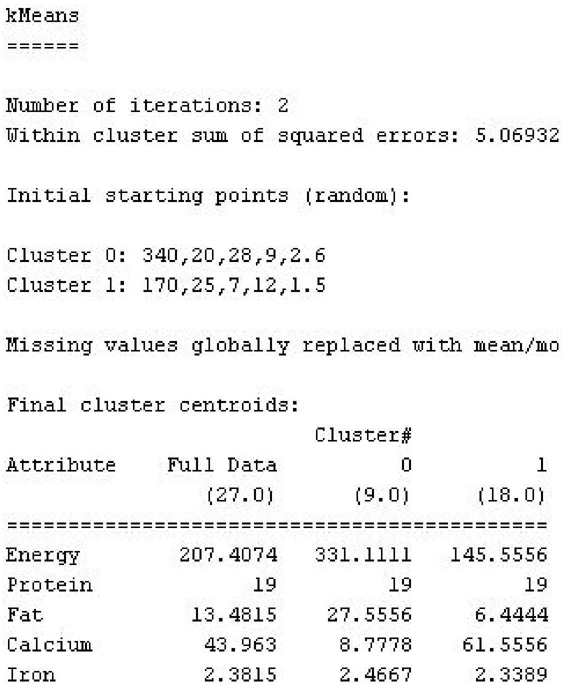
Instances: 27

Attributes: 6 Energy Protein Fat Calcium Iron

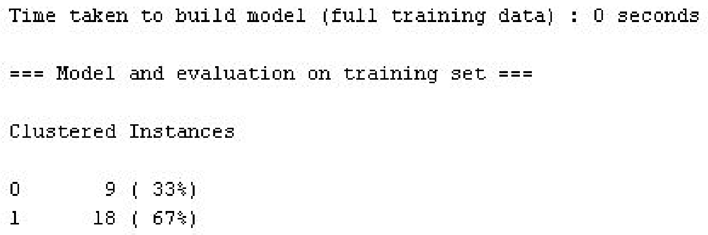
Ignored: Name

Test mode: evaluate on training data

Clustering model (full training set)



```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 5.069321339929419

Initial starting points (random):

Cluster 0: 340,20,28,9,2.6
Cluster 1: 170,25,7,12,1.5

Missing values globally replaced with mean/mode

Final cluster centroids:
                            Cluster#
Attribute    Full Data          0          1
                (27.0)      (9.0)     (18.0)
============================================
Energy        207.4074   331.1111   145.5556
Protein             19         19         19
Fat            13.4815    27.5556     6.4444
Calcium         43.963     8.7778    61.5556
Iron            2.3815     2.4667     2.3389
```



```
Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       9 ( 33%)
1      18 ( 67%)
```

When Number of cluster(K) = 5 with seed 10

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A “weka.core.EuclideanDistance -R first-last” -I 500 -num-slots 1 -S 10

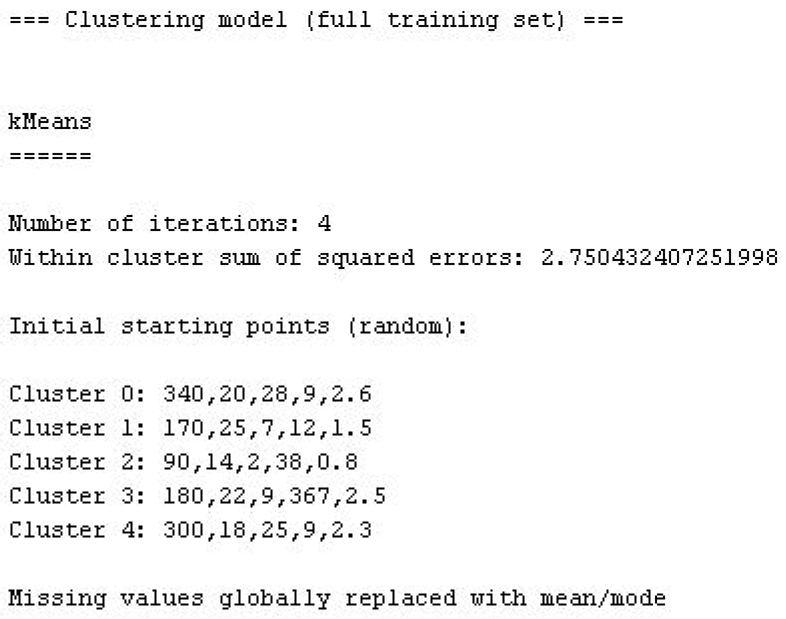
Relation: food

Instances: 27

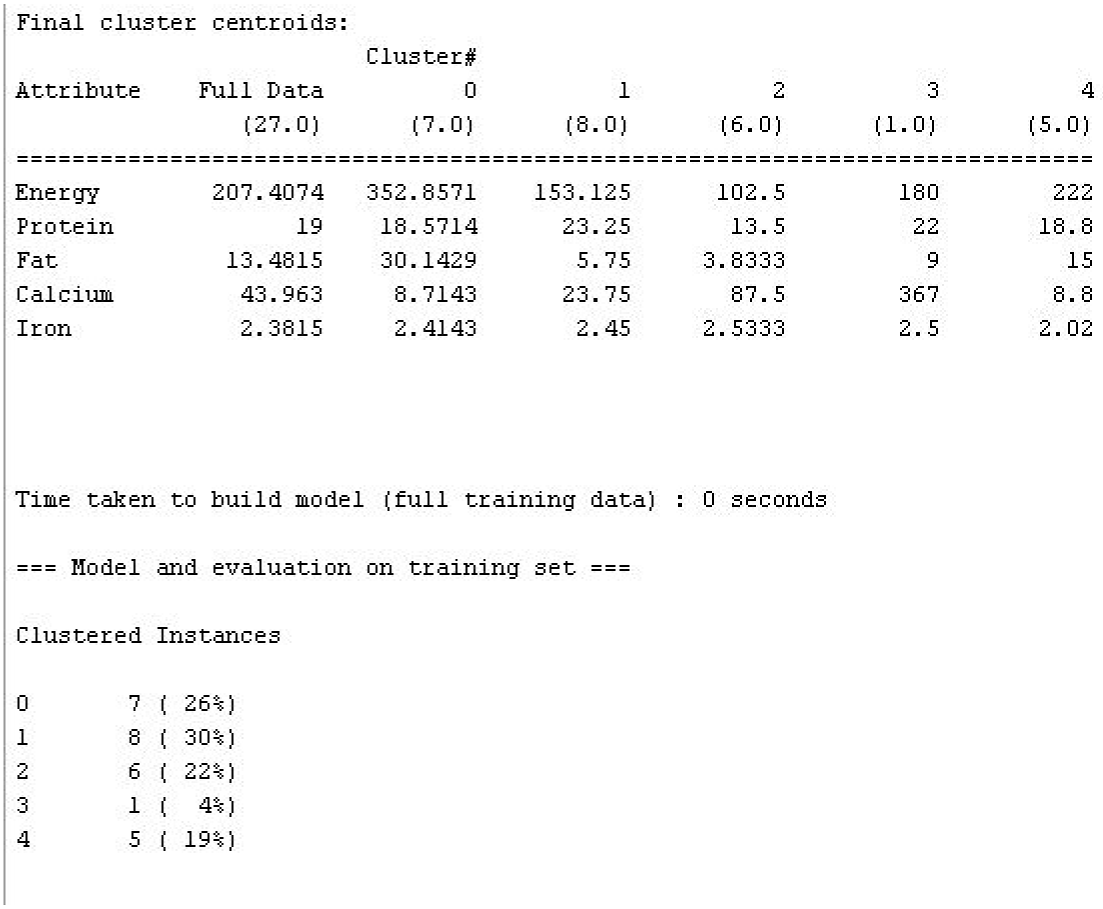
Attributes: 6 Energy Protein Fat Calcium Iron

Ignored: Name

Test mode: evaluate on training data



```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 2.750432407251998

Initial starting points (random):

Cluster 0: 340,20,28,9,2.6
Cluster 1: 170,25,7,12,1.5
Cluster 2: 90,14,2,38,0.8
Cluster 3: 180,22,9,367,2.5
Cluster 4: 300,18,25,9,2.3

Missing values globally replaced with mean/mode
```



```
Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2          3          4
                (27.0)      (7.0)      (8.0)      (6.0)      (1.0)      (5.0)
==============================================================================
Energy        207.4074   352.8571    153.125      102.5        180        222
Protein             19    18.5714      23.25       13.5         22       18.8
Fat            13.4815    30.1429       5.75     3.8333          9         15
Calcium         43.963     8.7143      23.75       87.5        367        8.8
Iron            2.3815     2.4143       2.45     2.5333        2.5       2.02




Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       7 ( 26%)
1       8 ( 30%)
2       6 ( 22%)
3       1 (  4%)
4       5 ( 19%)
```

1.3: Then try with a different seed value, i.e. different initial cluster centers. Compare the results with the previous results. Explain what the seed value controls.

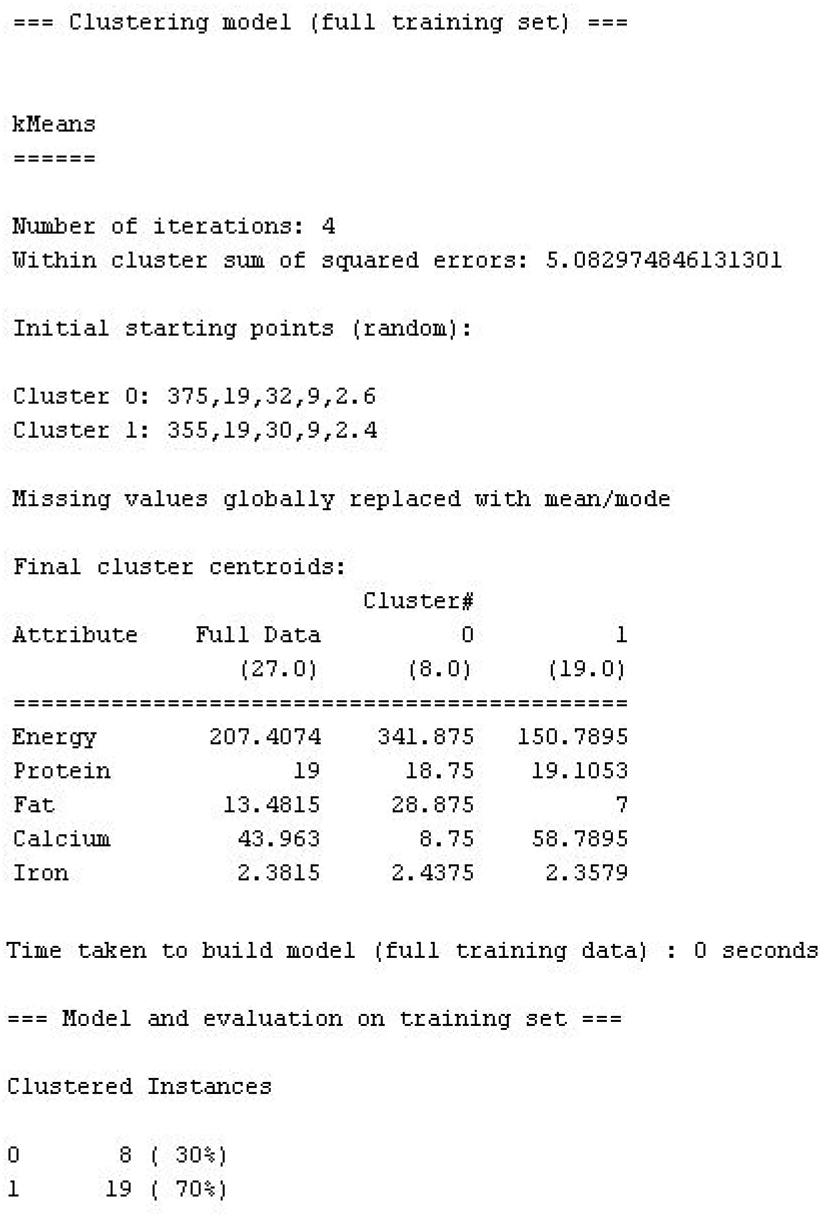
Changing seed value to 15 and recalculating.

At k=2 with seed at 15

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A “weka.core.EuclideanDistance -R first-last” -I 500 -num-slots 1 -S 15

Relation: food Instances: 27 Attributes: 6 Energy Protein Fat Calcium Iron Ignored: Name Test mode: evaluate on training data

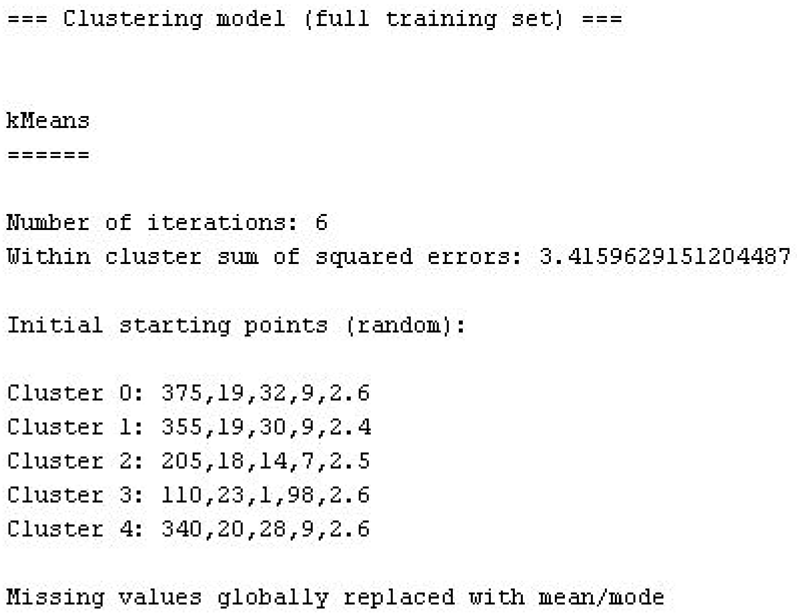


```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 5.082974846131301

Initial starting points (random):

Cluster 0: 375,19,32,9,2.6
Cluster 1: 355,19,30,9,2.4

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                (27.0)      (8.0)     (19.0)
============================================
Energy        207.4074    341.875   150.7895
Protein             19      18.75    19.1053
Fat            13.4815     28.875          7
Calcium         43.963       8.75    58.7895
Iron            2.3815     2.4375     2.3579

Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       8 ( 30%)
1      19 ( 70%)
```

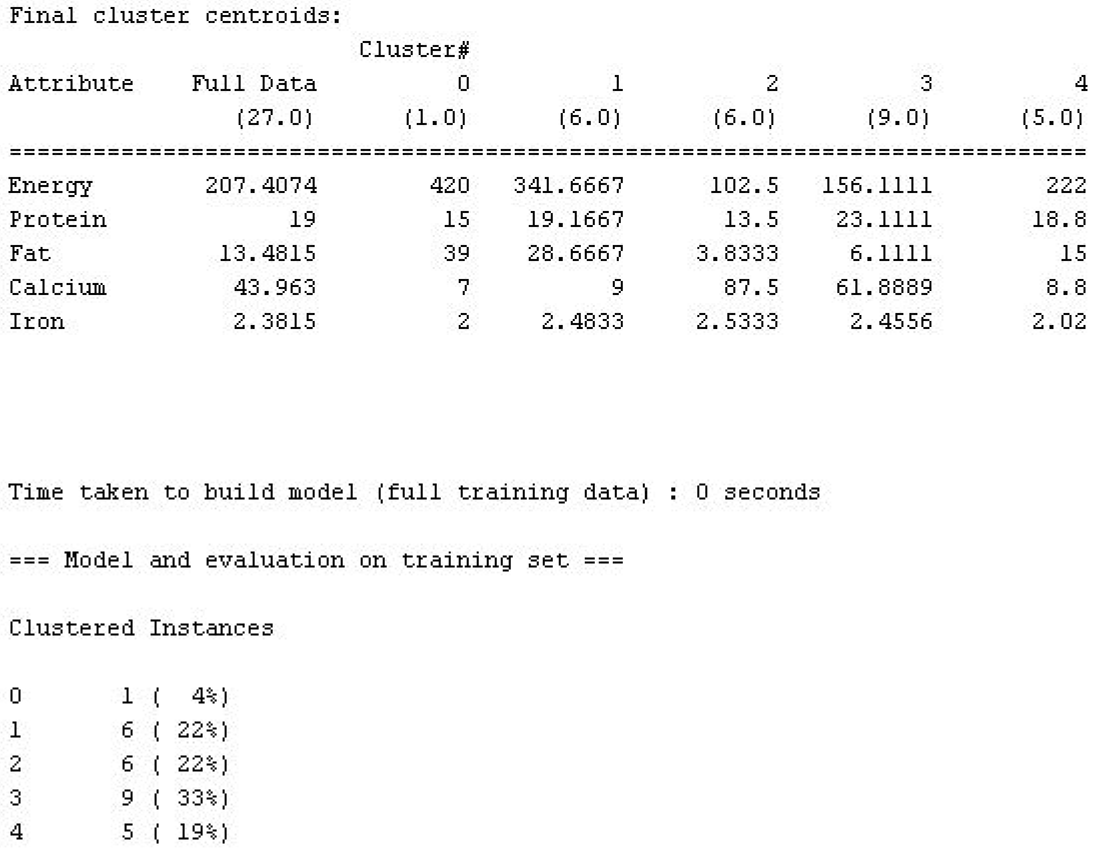
At k = 5 with seed at 15

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A “weka.core.EuclideanDistance -R first-last” -I 500 -num-slots 1 -S 15 Relation: food Instances: 27 Attributes: 6 Energy Protein Fat Calcium Iron Ignored: Name Test mode: evaluate on training data



```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 3.4159629151204487

Initial starting points (random):

Cluster 0: 375,19,32,9,2.6
Cluster 1: 355,19,30,9,2.4
Cluster 2: 205,18,14,7,2.5
Cluster 3: 110,23,1,98,2.6
Cluster 4: 340,20,28,9,2.6

Missing values globally replaced with mean/mode
```



```
Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1          2          3          4
                (27.0)      (1.0)      (6.0)      (6.0)      (9.0)      (5.0)
==============================================================================
Energy        207.4074        420   341.6667      102.5   156.1111        222
Protein             19         15    19.1667       13.5    23.1111       18.8
Fat            13.4815         39    28.6667     3.8333     6.1111         15
Calcium         43.963          7          9       87.5    61.8889        8.8
Iron            2.3815          2     2.4833     2.5333     2.4556       2.02




Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       1 (  4%)
1       6 ( 22%)
2       6 ( 22%)
3       9 ( 33%)
4       5 ( 19%)
```

Impact of changed seed value

The seed value is used to randomly generate the initial k number of means, which are used as initial centroids of clusters, and initial clustering is done using them.

While checking the impact of seed, it can be seen that in the case of *k=5*, with different seed values of 10 and 15, it take more iterations before classifying and along with increased in sum of square error(SSE).

```
## Table of Comparison with Number of Cluster set to 5
##            Iteration  SSE
## Seed = 10          4 2.75
## Seed = 15          6 3.41
```

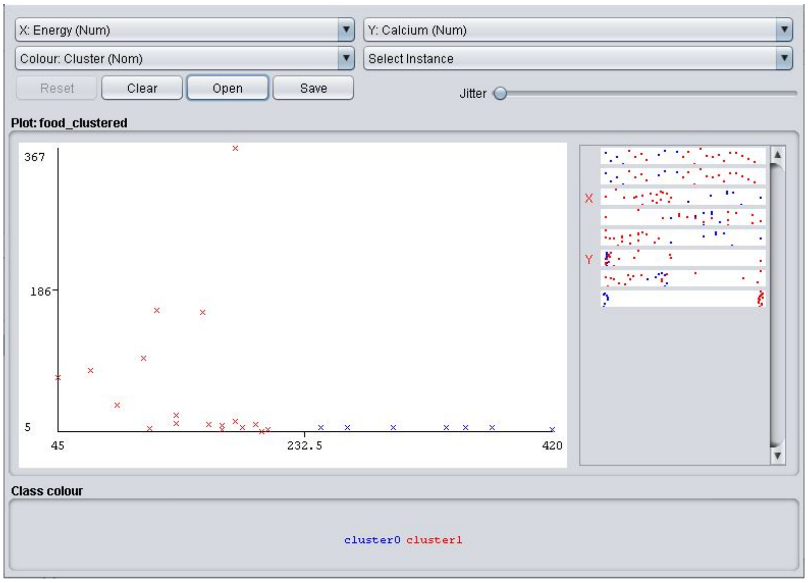
```
## Table of Comparison with Number of Cluster set to 2
```

```
##            Iteration  SSE
## Seed = 10          2 5.06
## Seed = 15          4 5.08
```

1.4: Do you think the clusters are “good” clusters? (Are all of its members “similar” to each other? Are members from different clusters dissimilar?)

When k=2 regardless of seed value, it can be seen that classification is majorly done in two variables, namely Energy and Calcium, while reaming are just randomly assigned to clusters based on centroid mean.

When k=5, it can be seen that clusters majorly formed between 3 Variable Energy, Fat and Calcium yet no clear information can be derived from it thus these many clusters for such small dataset do not provide much of information in order to make a decision thus cannot be considered an as good cluster.



X: Energy (Num)

Y: Calcium (Num)

Colour: Cluster (Nom)

Select Instance

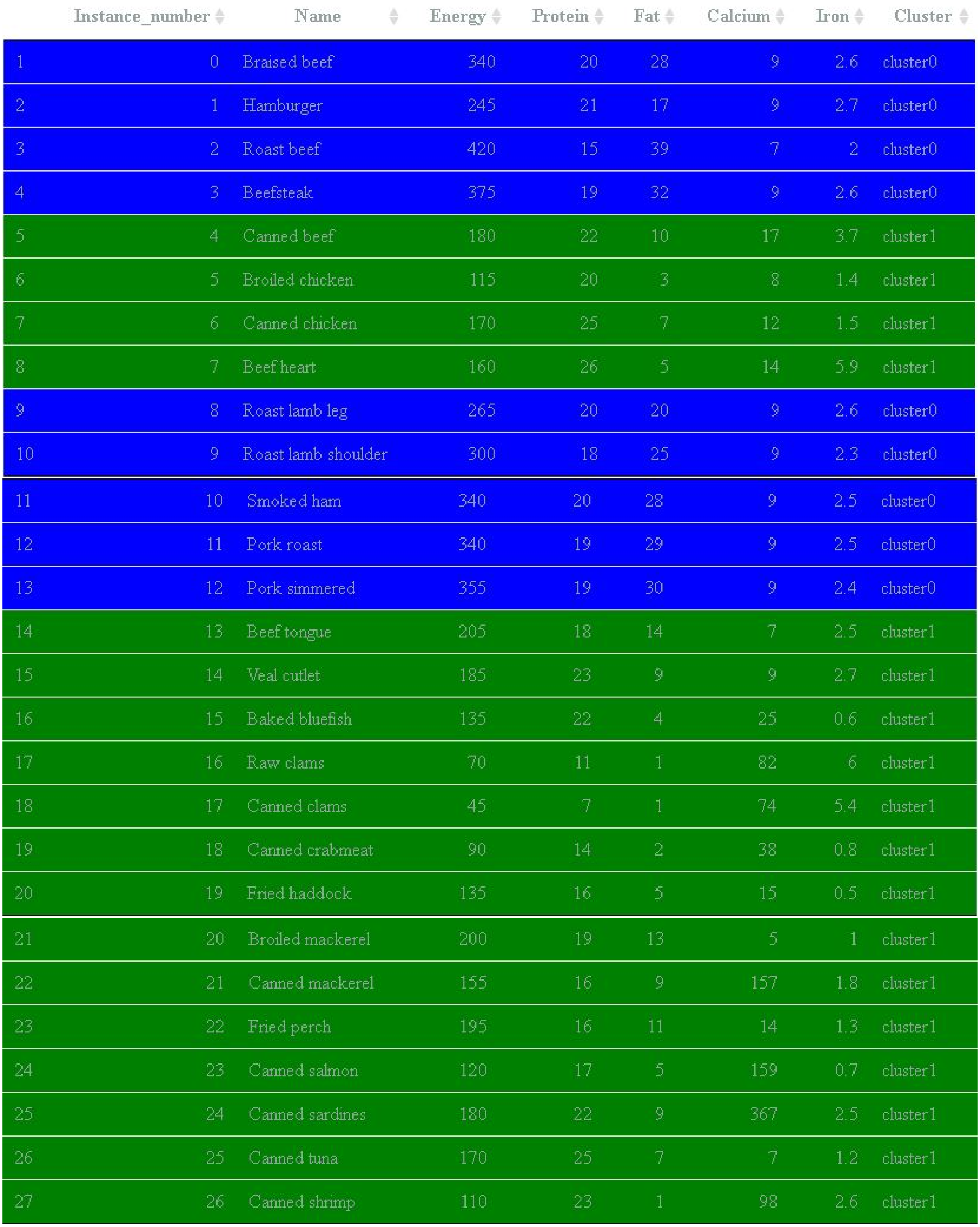
Plot: food_clustered

Class colour

```
## Cluster when k=2 between Energy and Calcium
```

1.5: What does each cluster represent? Choose one of the results. Make up labels (words or phrases in English) which characterize each cluster

When K=2 and seed is 10, we could consider the division of data into based on centroid namely as Cluster-0 as **High Energy - Low Calcium** which is marked with blue colour and Cluster-1 as **Low Energy - High Calcium** the cluster which is marked with green colour.



	Instance_number	Name	Energy	Protein	Fat	Calcium	Iron	Cluster
1	0	Braised beef	340	20	28	9	2.6	cluster0
2	1	Hamburger	245	21	17	9	2.7	cluster0
3	2	Roast beef	420	15	39	7	2	cluster0
4	3	Beefsteak	375	19	32	9	2.6	cluster0
5	4	Canned beef	180	22	10	17	3.7	cluster1
6	5	Broiled chicken	115	20	3	8	1.4	cluster1
7	6	Canned chicken	170	25	7	12	1.5	cluster1
8	7	Beef heart	160	26	5	14	5.9	cluster1
9	8	Roast lamb leg	265	20	20	9	2.6	cluster0
10	9	Roast lamb shoulder	300	18	25	9	2.3	cluster0
11	10	Smoked ham	340	20	28	9	2.5	cluster0
12	11	Pork roast	340	19	29	9	2.5	cluster0
13	12	Pork simmered	355	19	30	9	2.4	cluster0
14	13	Beef tongue	205	18	14	7	2.5	cluster1
15	14	Veal cutlet	185	23	9	9	2.7	cluster1
16	15	Baked bluefish	135	22	4	25	0.6	cluster1
17	16	Raw clams	70	11	1	82	6	cluster1
18	17	Canned clams	45	7	1	74	5.4	cluster1
19	18	Canned crabmeat	90	14	2	38	0.8	cluster1
20	19	Fried haddock	135	16	5	15	0.5	cluster1
21	20	Broiled mackerel	200	19	13	5	1	cluster1
22	21	Canned mackerel	155	16	9	157	1.8	cluster1
23	22	Fried perch	195	16	11	14	1.3	cluster1
24	23	Canned salmon	120	17	5	159	0.7	cluster1
25	24	Canned sardines	180	22	9	367	2.5	cluster1
26	25	Canned tuna	170	25	7	7	1.2	cluster1
27	26	Canned shrimp	110	23	1	98	2.6	cluster1

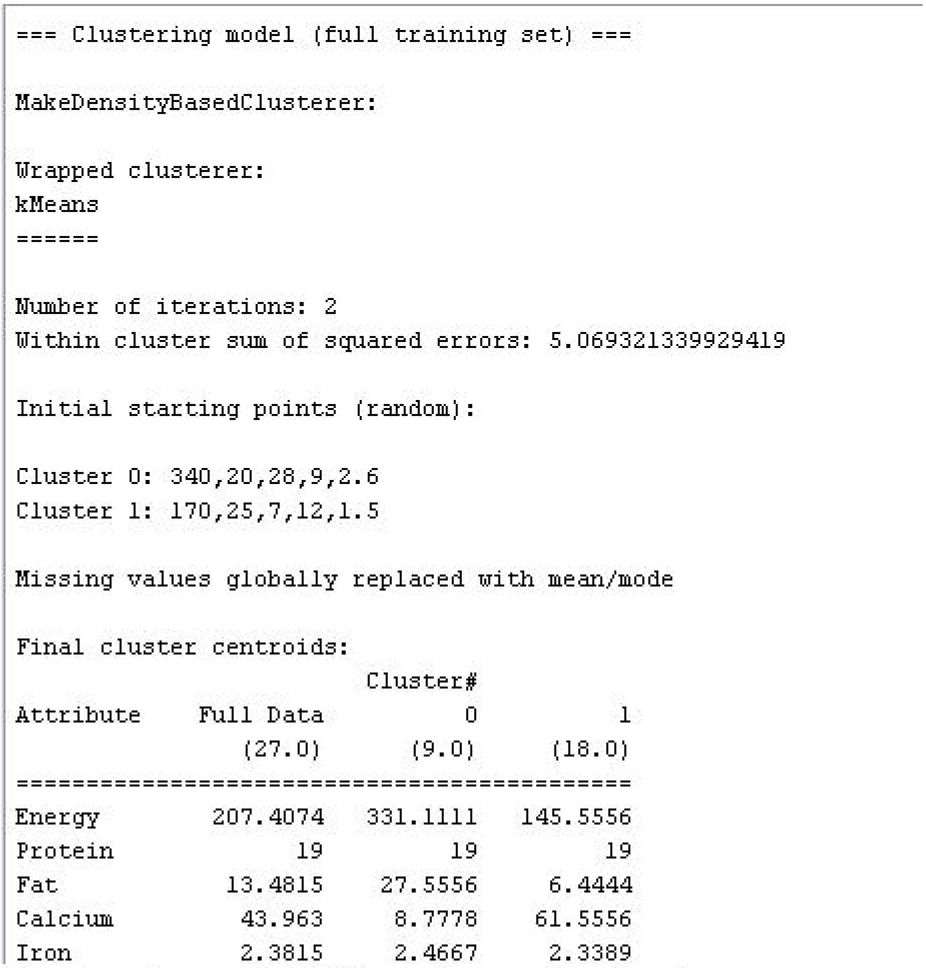
Question 2 : MakeDensityBasedClusters

Use the SimpleKMeans clusterer which gave the result you haven chosen in 5 Experiment with at least two different standard deviations. Compare the results.

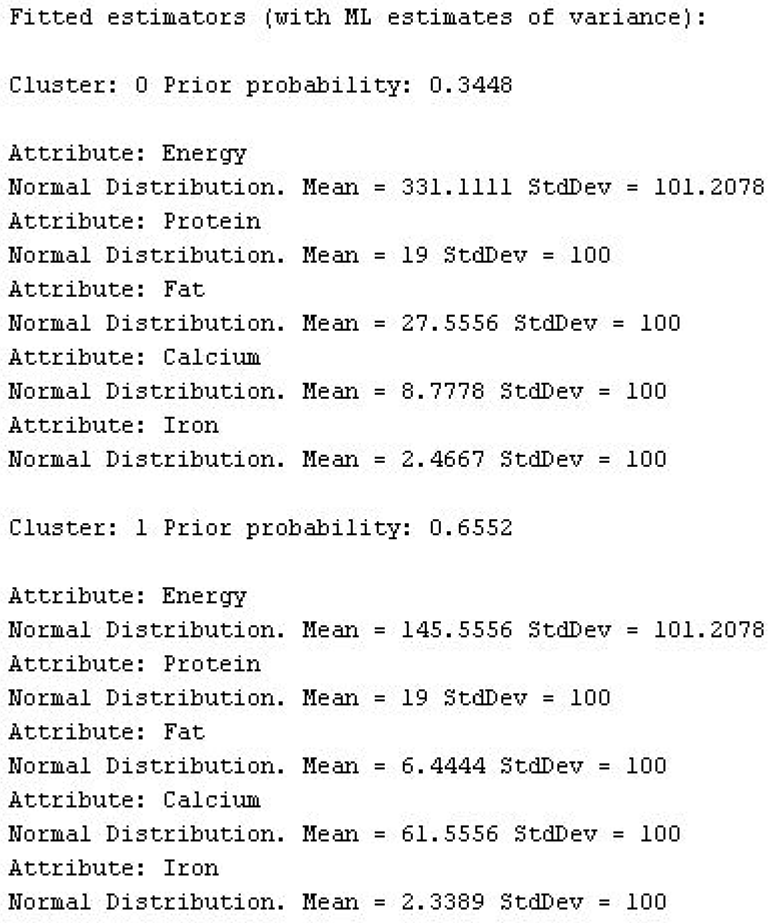
We are using SimpleKMeans with 2 clusters and seed as 10 to make DensityBased Cluster with a different value of standard deviations. Over here, we have change minimal standard deviation, which allows us to set a minimum threshold for standard deviation applied to the normal distribution of features in all dimensions.

Case : When minStdDev = 100

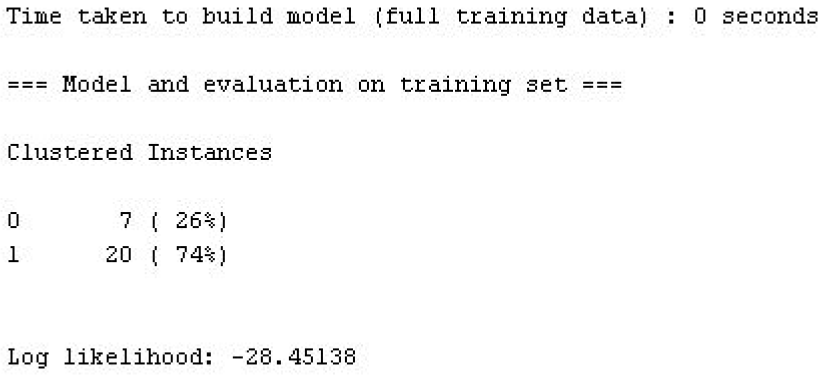
```
=== Run information ===

Scheme: weka.clusterers.MakeDensityBasedClusterer -M 100.0 -W weka.clusterers.SimpleKMeans – -init 0 -
max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A
“weka.core.EuclideanDistance -R first-last” -I 500 -num-slots 1 -S 10

Relation: food

Instances: 27

Attributes: 6 Energy Protein Fat Calcium Iron

Ignored: Name

Test mode: evaluate on training data
```



```
=== Clustering model (full training set) ===

MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 5.069321339929419

Initial starting points (random):

Cluster 0: 340,20,28,9,2.6
Cluster 1: 170,25,7,12,1.5

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                (27.0)      (9.0)     (18.0)
============================================
Energy        207.4074   331.1111   145.5556
Protein             19         19         19
Fat            13.4815    27.5556     6.4444
Calcium         43.963     8.7778    61.5556
Iron            2.3815     2.4667     2.3389
```



```
Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.3448

Attribute: Energy
Normal Distribution. Mean = 331.1111 StdDev = 101.2078
Attribute: Protein
Normal Distribution. Mean = 19 StdDev = 100
Attribute: Fat
Normal Distribution. Mean = 27.5556 StdDev = 100
Attribute: Calcium
Normal Distribution. Mean = 8.7778 StdDev = 100
Attribute: Iron
Normal Distribution. Mean = 2.4667 StdDev = 100

Cluster: 1 Prior probability: 0.6552

Attribute: Energy
Normal Distribution. Mean = 145.5556 StdDev = 101.2078
Attribute: Protein
Normal Distribution. Mean = 19 StdDev = 100
Attribute: Fat
Normal Distribution. Mean = 6.4444 StdDev = 100
Attribute: Calcium
Normal Distribution. Mean = 61.5556 StdDev = 100
Attribute: Iron
Normal Distribution. Mean = 2.3389 StdDev = 100
```



```
Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0       7 ( 26%)
1      20 ( 74%)


Log likelihood: -28.45138
```

Case : When minStdDev = 0.001

=== Run information ===

Scheme: weka.clusterers.MakeDensityBasedClusterer -M 0.001 -W weka.clusterers.SimpleKMeans – -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A “weka.core.EuclideanDistance -R first-last” -I 500 -num-slots 1 -S 10

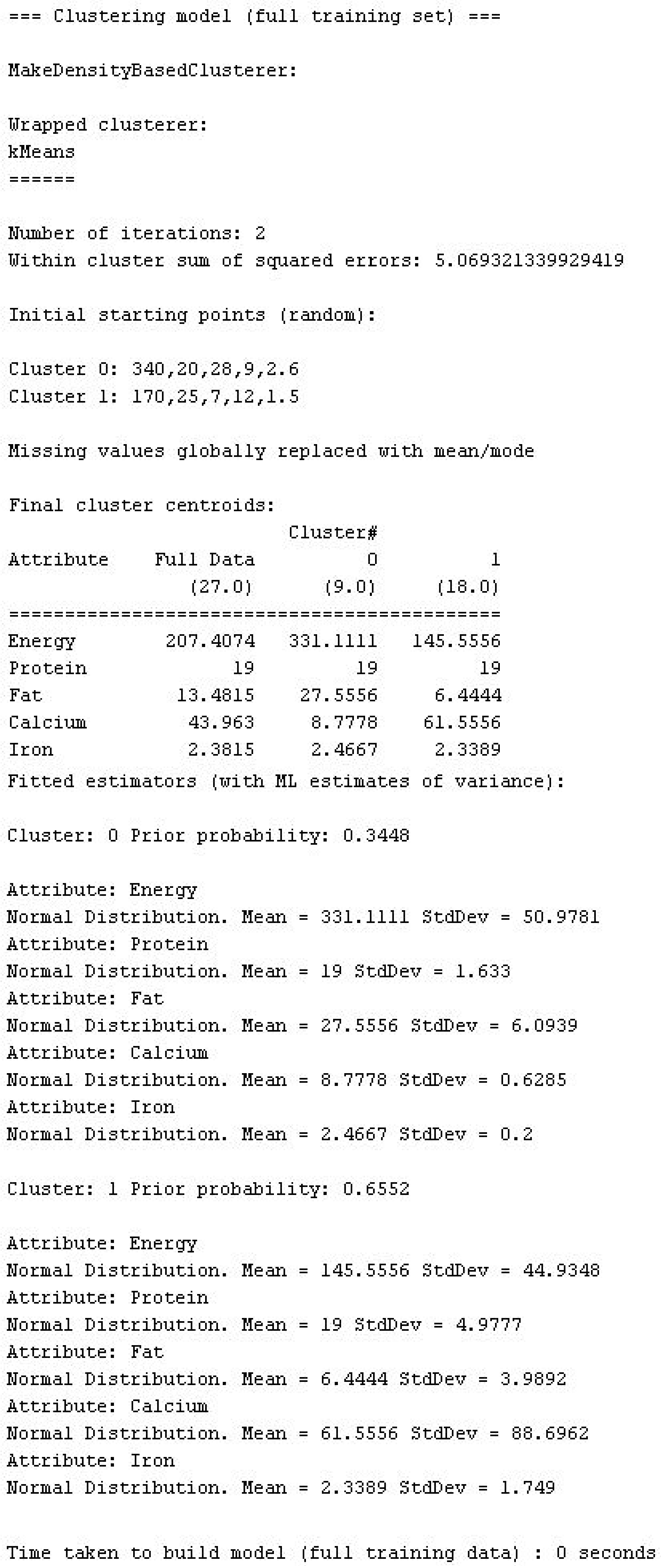
Relation: food

Instances: 27

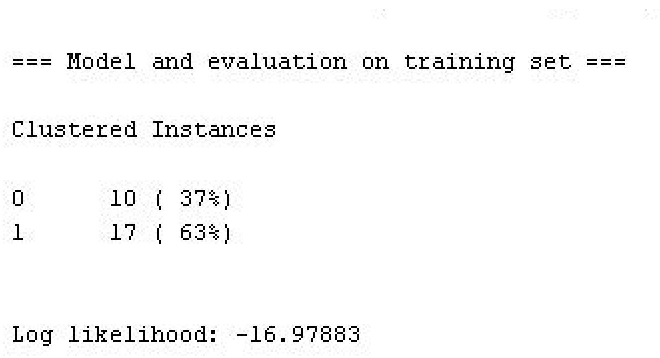
Attributes: 6 Energy Protein Fat Calcium Iron

Ignored: Name

Test mode: evaluate on training data



```
=== Clustering model (full training set) ===

MakeDensityBasedClusterer:

Wrapped clusterer:
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 5.069321339929419

Initial starting points (random):

Cluster 0: 340,20,28,9,2.6
Cluster 1: 170,25,7,12,1.5

Missing values globally replaced with mean/mode

Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                (27.0)      (9.0)     (18.0)
============================================
Energy        207.4074   331.1111   145.5556
Protein             19         19         19
Fat            13.4815    27.5556     6.4444
Calcium         43.963     8.7778    61.5556
Iron            2.3815     2.4667     2.3389
Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.3448

Attribute: Energy
Normal Distribution. Mean = 331.1111 StdDev = 50.9781
Attribute: Protein
Normal Distribution. Mean = 19 StdDev = 1.633
Attribute: Fat
Normal Distribution. Mean = 27.5556 StdDev = 6.0939
Attribute: Calcium
Normal Distribution. Mean = 8.7778 StdDev = 0.6285
Attribute: Iron
Normal Distribution. Mean = 2.4667 StdDev = 0.2

Cluster: 1 Prior probability: 0.6552

Attribute: Energy
Normal Distribution. Mean = 145.5556 StdDev = 44.9348
Attribute: Protein
Normal Distribution. Mean = 19 StdDev = 4.9777
Attribute: Fat
Normal Distribution. Mean = 6.4444 StdDev = 3.9892
Attribute: Calcium
Normal Distribution. Mean = 61.5556 StdDev = 88.6962
Attribute: Iron
Normal Distribution. Mean = 2.3389 StdDev = 1.749


Time taken to build model (full training data) : 0 seconds
```



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

Clustered Instances

0       10 ( 37%)
1       17 ( 63%)


Log likelihood: -16.97883
```

As we can see, when the standard deviation changes from 0.001 to 100 variation in cluster increase. Thus when density-based clustering is applied cluster which was part of Cluster 0 were moved into Cluster 1, as now because of changed variation density reachable point changed and thus new clusters with different numbers were obtained.