Lab Report: Lab1 - Clustering Lab

Dhanush Kumar Reddy Narayana Reddy (dhana004), Udaya Shanker Mohanan Nair (udamo524)

2025-03-12

Introduction

Implementation of Kmeans, SimpleKMeans and density-based method.

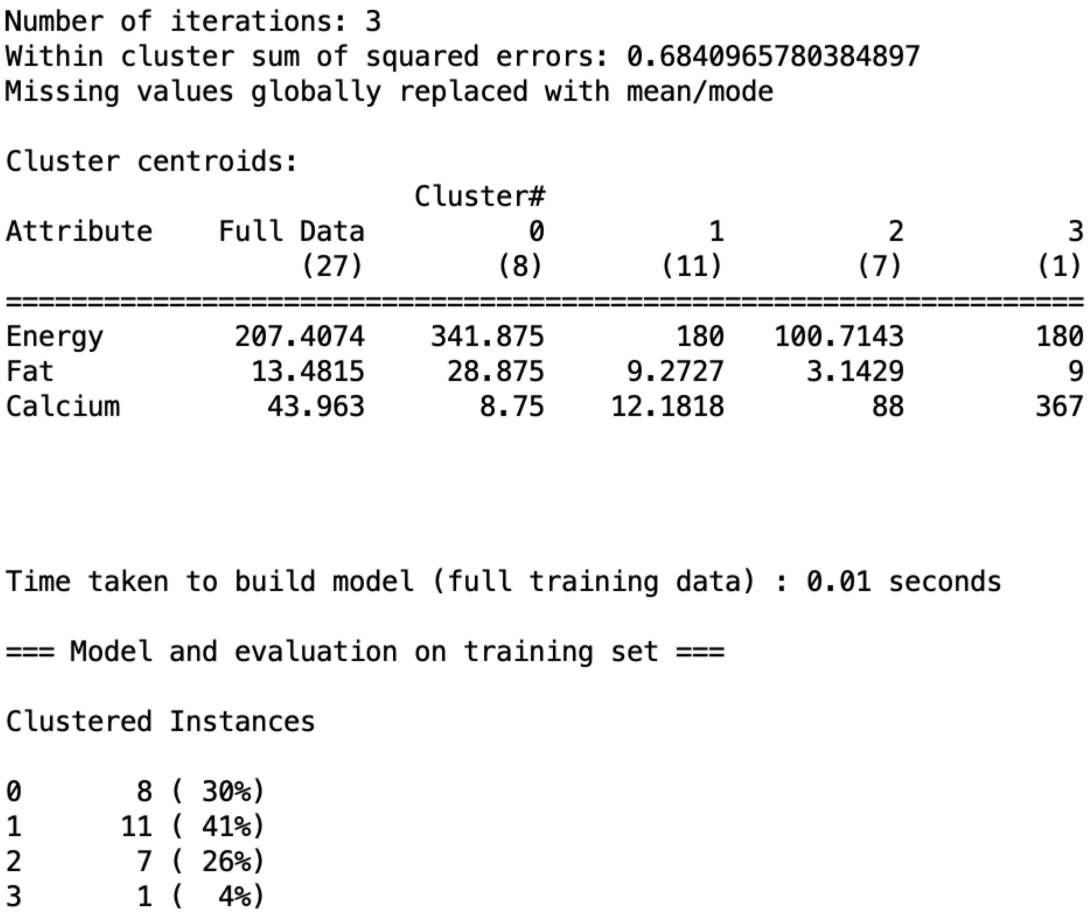
SimpleKMeans

1. Choose a set of attributes for clustering and give a motivation. (Hint: always ignore attribute "name". Why does the name attribute need to be ignored?)

We choose attributes energy, fat, calcium as viewing values directly we could understand that energy, protein and fat contribute more than other attributes given. Name is always ignored because it is an identical attribute and does not contribute during clustering.

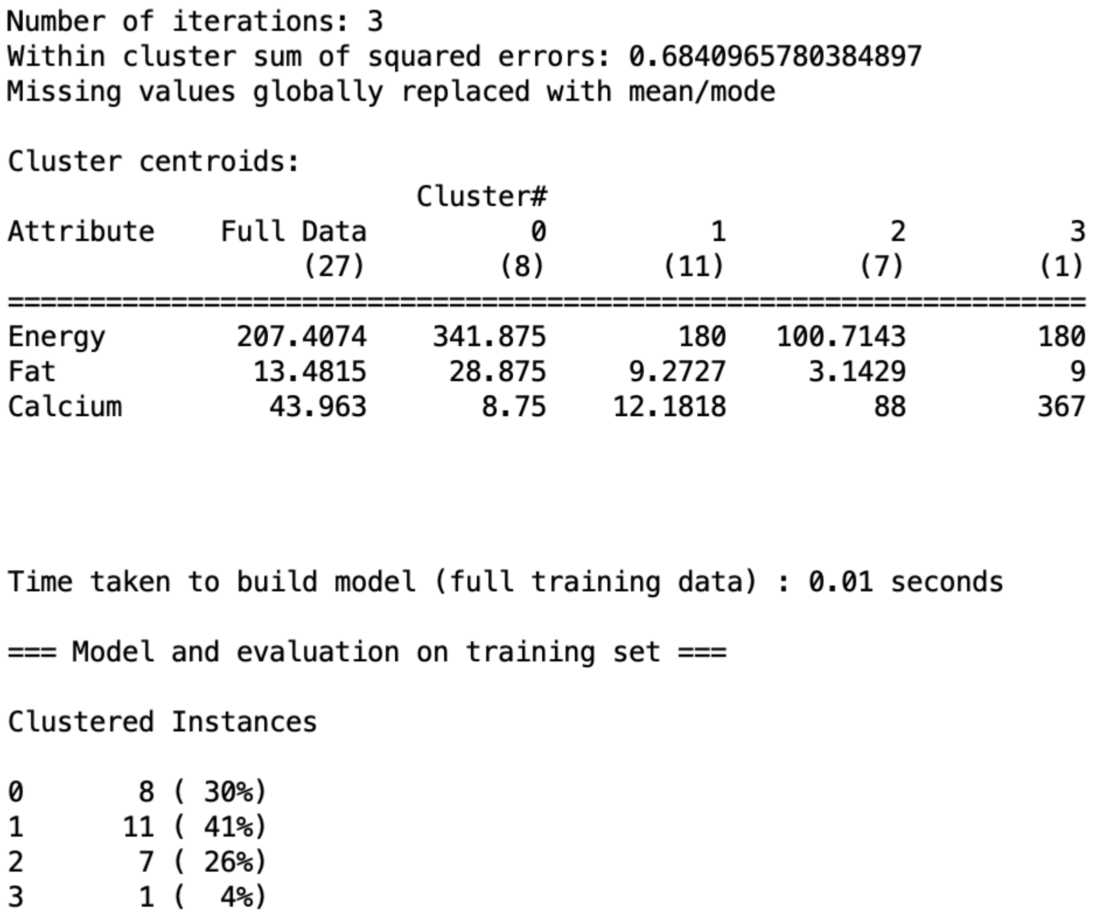
2. Experiment with at least two different numbers of clusters, e.g. 2 and 5, but with the same seed value 10.

In first experiment, tried with cluster value as 2 and seed value as 10 for all attributes mentioned in the first questions answer.



```
Number of iterations: 3
Within cluster sum of squared errors: 0.6840965780384897
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute    Full Data           0           1           2           3
                  (27)         (8)        (11)         (7)         (1)
======================================================================
Energy        207.4074     341.875         180    100.7143         180
Fat            13.4815      28.875      9.2727      3.1429           9
Calcium         43.963        8.75     12.1818          88         367




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

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

Clustered Instances

0       8 ( 30%)
1      11 ( 41%)
2       7 ( 26%)
3       1 (  4%)
```

In second experiment, tried with cluster value as 4 and seed value as 10 for all attributes mentioned in the first questions answer.

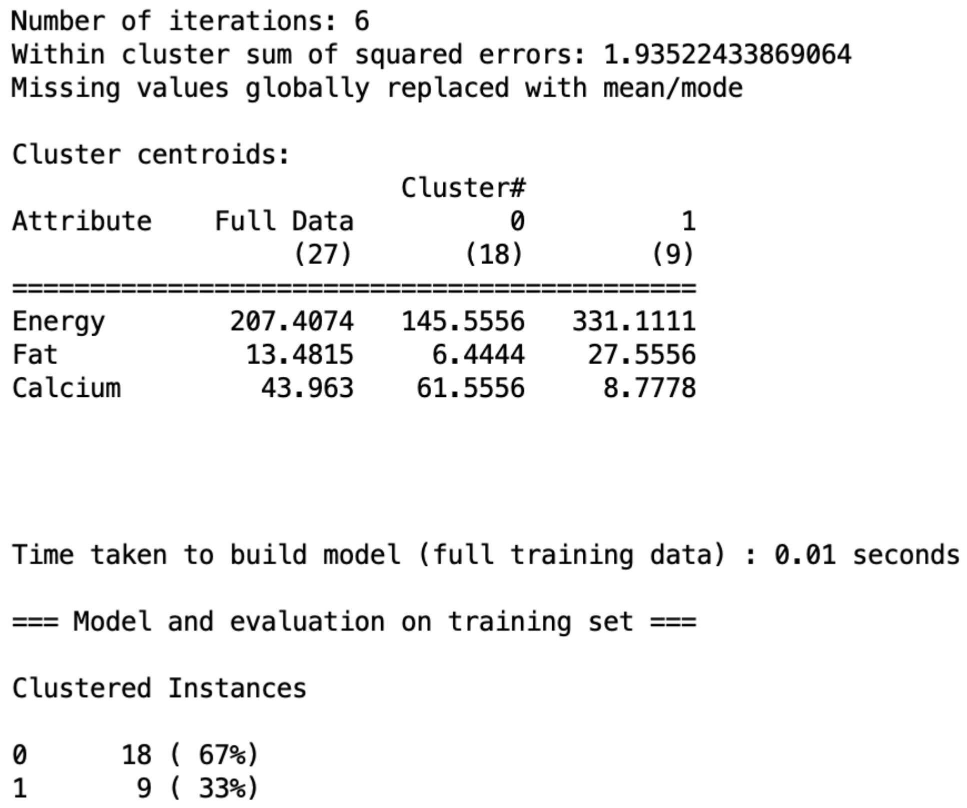


```
Number of iterations: 3
Within cluster sum of squared errors: 0.6840965780384897
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute    Full Data             0            1            2            3
                  (27)           (8)         (11)          (7)          (1)
==================================================================
Energy        207.4074       341.875          180     100.7143          180
Fat            13.4815        28.875       9.2727       3.1429            9
Calcium         43.963          8.75      12.1818           88          367




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

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

Clustered Instances

0       8 ( 30%)
1      11 ( 41%)
2       7 ( 26%)
3       1 (  4%)
```

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.

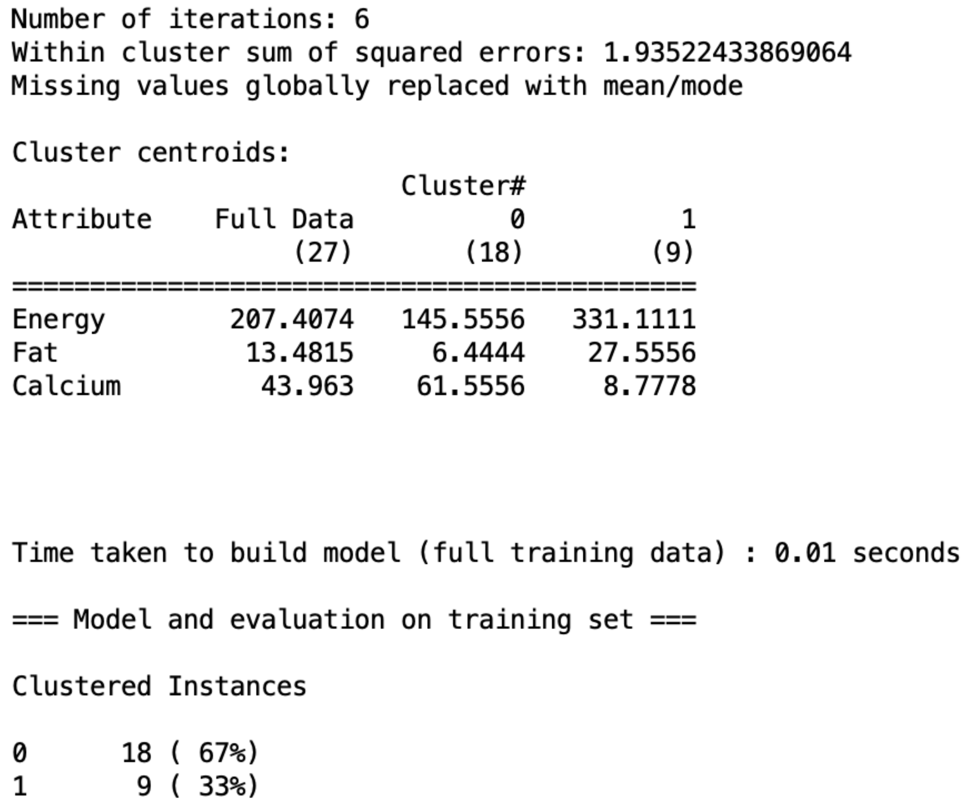
Changing seed value to 20.

Results of first experiment



```
Number of iterations: 6
Within cluster sum of squared errors: 1.93522433869064
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                  (27)       (18)        (9)
============================================
Energy        207.4074   145.5556   331.1111
Fat            13.4815     6.4444    27.5556
Calcium         43.963    61.5556     8.7778




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

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

Clustered Instances

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

Results of second experiment



```
Number of iterations: 6
Within cluster sum of squared errors: 1.93522433869064
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
                  (27)       (18)        (9)
============================================
Energy        207.4074   145.5556   331.1111
Fat            13.4815     6.4444    27.5556
Calcium         43.963    61.5556     8.7778




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

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

Clustered Instances

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

Comparison of results

For Cluster 2

Cluster sum of squared error changed from 1.94879 -> 1.93522.

For Cluster 4

Cluster sum of squared error changed from 0.6841 -> 0.7243.

Seed value controls the initial cluster centers, which controls the clustering.Different seed value results in different cluster formation and squeared error values.

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

Clusters are not properly good, as for k=2, the clustering is weak where one cluster dominates the other for both cases of seed(10 or 20). When k=4, the clustering is better and clusters are distributed more fairly when compared to other k value for both the seed values.

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

Below table shows the cluster distribution when k=2 and seed value is 10.

Table 1: Cluster Labelling

Cluster	Energy	Fat	Calcium	Label
0	341.8750	28.8750	8.7500	High-Fat, Energy, calicum product
1	180.0000	9.2727	12.1818	Balanced Fat Energy Calcium product
2	100.7143	3.1429	88.0000	Rich Calcium Dairy Product

Cluster	Energy	Fat	Calcium	Label
3	180.0000	9.0000	367.0000	High Calcium Product