ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN

KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG
Bài tập Thực hành 4

CLUSTERING

Nhóm thực hiện

1. Hồng Thanh Hoài 1612855
2. Huỳnh Minh Huấn 1612858

Giáo viên lý thuyết
PGS.TS Lê Hoài Bắc

Giáo viên hướng dẫn
Nguyễn Ngọc Thảo

Tháng 06 năm 2019

Lời cảm ơn

Trong quá trình thực hiện bài tập này, nhóm chúng em đã nhận được rất nhiều sự giúp đỡ cũng như hỗ trợ từ các thầy cô Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM và các bạn bè trong trường. Nhóm chúng em xin bày tỏ lòng cảm ơn chân thành đến mọi người vì đã hướng dẫn, chỉ bảo rất tận tình.

Đặc biệt, nhóm chúng em xin bày tỏ lòng biết ơn sâu sắc đến các thầy cô khoa Công nghệ thông tin, cụ thể hơn là thầy Lê Hoài Bắc đã giảng dạy rất kĩ lưỡng để chúng em nắm rõ kiến thức và cô Nguyễn Ngọc Thảo đã hướng dẫn chúng em thực hiện bài tập này rất nhiệt tình.

Một lần nữa, chúng em xin bày tỏ lòng biết ơn sâu sắc đến với các thầy cô và bạn bè.

Tháng 06 năm 2019,
Đại học Khoa học Tự nhiên, ĐHQG-HCM.

Mục lục

1 Giới thiệu nhóm và phân công công việc

1.1 Giới thiệu nhóm

Nhóm gồm 2 thành viên.

STT	Họ và tên	MSSV	Email	SĐT
1	Hồng Thanh Hoài	1612855	hthoai1006@gmail.com	0965596807
2	Huỳnh Minh Huấn	1612858	minhhuanhuynh289@gmail.com	0824540646

1.2 Phân công công việc

STT	Họ và tên	Công việc
1	Hồng Thanh Hoài	Phần Weka.
2	Huỳnh Minh Huấn	Phần cài đặt.

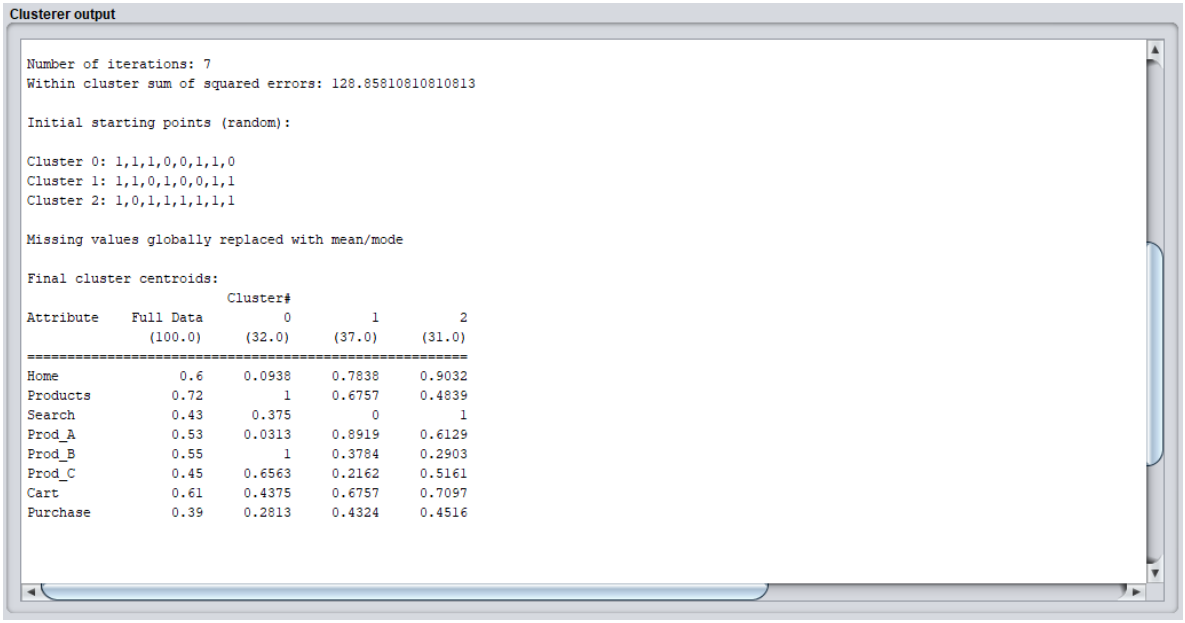
2 Nội dung

2.1 Weka

2.1.1 Câu 1

Bảng 1: Bảng với các giá trị k từ 3–8.

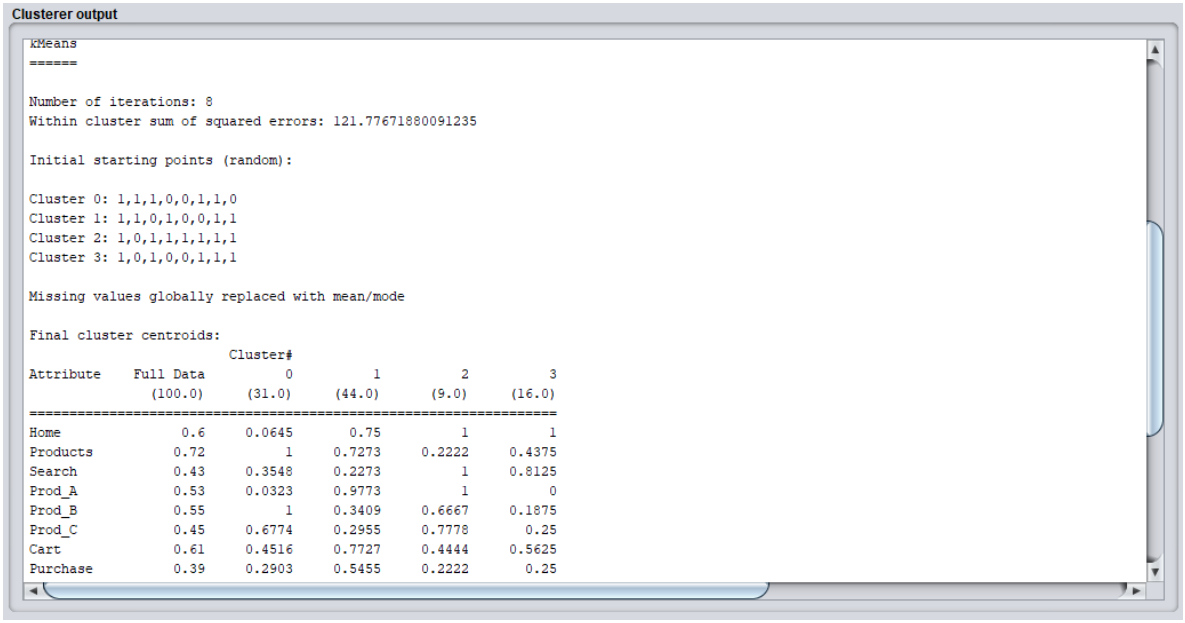
k	SSE	Cluster centroids							
			Home	Products	Search	Prod_A	Prod_B	Prod_C	Cart	Purchase
3	128.8581	1	0.0938	1	0.375	0.0313	1	0.6563	0.4375	0.2813
		2	0.7838	0.6757	0	0.8919	0.3784	0.2162	0.6757	0.4324
		3	0.9032	0.4839	1	0.6129	0.2903	0.5161	0.7097	0.4516
4	121.7767	1	0.0645	1	0.3548	0.0323	1	0.6774	0.4516	0.2903
		2	0.75	0.7273	0.2273	0.9773	0.3409	0.2955	0.7727	0.5455
		3	1	0.2222	1	1	0.6667	0.7778	0.4444	0.2222
		4	1	0.4375	0.8125	0	0.1875	0.25	0.5625	0.25
5	113.5826	1	0.9615	0.6923	0.6538	0.4615	0.3846	0.5385	0.4615	0
		2	0.6667	0.6667	0	0.963	0.4444	0	0.6296	0.5185
		3	1	0	1	1	0.8	0.8	0.8	0.4
		4	0.8571	0.5714	0.8571	0.7143	0.0714	0.5714	1	1
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
6	109.3612	1	0.931	0.7241	0.7241	0.5862	0.3448	0.6552	0.5172	0.1379
		2	0.9583	0.875	0.083	0.9167	0.3333	0.0833	0.7083	0.5833
		3	1	0	1	1	1	0.75	0.75	0.5
		4	1	0.1667	1	0.1667	0	0.3333	1	0.8333
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1111	0.1111	1	0.5556	0	0.6667	0.5556
7	93.7901	1	0.9048	0.9048	0.7143	0.7619	0.381	0.9048	0.6667	0.1905
		2	1	0.9375	0.125	0.875	0.4375	0.125	1	0.8125
		3	1	0	1	1	1	0.75	0.75	0.5
		4	1	0.2857	1	0.1429	0	0.2857	1	0.7142
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1429	0.1429	1	0.2857	0	1	0.8571
		7	0.8235	0.4118	0.2941	0.6471	0.3529	0	0	0
8	88.9319	1	0.8889	1	0.6667	0.7778	0.3889	0.9444	0.6667	0.2222
		2	1	0.9375	0.125	0.875	0.4375	0.125	1	0.8125
		3	1	0	1	1	0.6667	0.8333	0.6667	0.3333
		4	1	0.5	1	0	0	1	1	1
		5	0	1	0.3214	0	1	0.6786	0.5	0.3214
		6	0	0.1429	0.1429	1	0.2857	0	1	0.8571
		7	0.75	0.5	0	0.8333	0.8333	0	0	0
		8	1	0.2727	1	0.1818	0.2727	0	0.5455	0.2727



Clusterer output

```
Number of iterations: 7
Within cluster sum of squared errors: 128.85810810810813

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2
               (100.0)     (32.0)     (37.0)     (31.0)
============================================================
Home                0.6     0.0938     0.7838     0.9032
Products           0.72          1     0.6757     0.4839
Search             0.43      0.375          0          1
Prod_A             0.53     0.0313     0.8919     0.6129
Prod_B             0.55          1     0.3784     0.2903
Prod_C             0.45     0.6563     0.2162     0.5161
Cart               0.61     0.4375     0.6757     0.7097
Purchase           0.39     0.2813     0.4324     0.4516
```

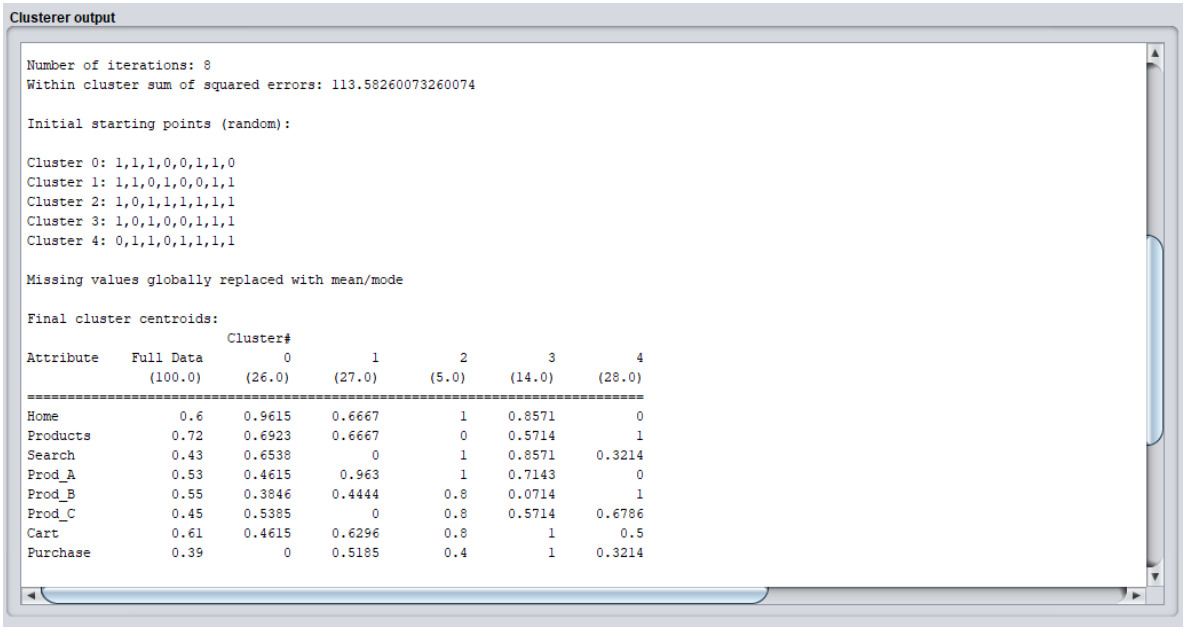
Hình 1: $k = 3$.



Clusterer output

```
kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 121.77671880091235

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3
               (100.0)     (31.0)     (44.0)      (9.0)     (16.0)
================================================================
Home                0.6     0.0645       0.75          1          1
Products           0.72          1     0.7273     0.2222     0.4375
Search             0.43     0.3548     0.2273          1     0.8125
Prod_A             0.53     0.0323     0.9773          1          0
Prod_B             0.55          1     0.3409     0.6667     0.1875
Prod_C             0.45     0.6774     0.2955     0.7778       0.25
Cart               0.61     0.4516     0.7727     0.4444     0.5625
Purchase           0.39     0.2903     0.5455     0.2222       0.25
```

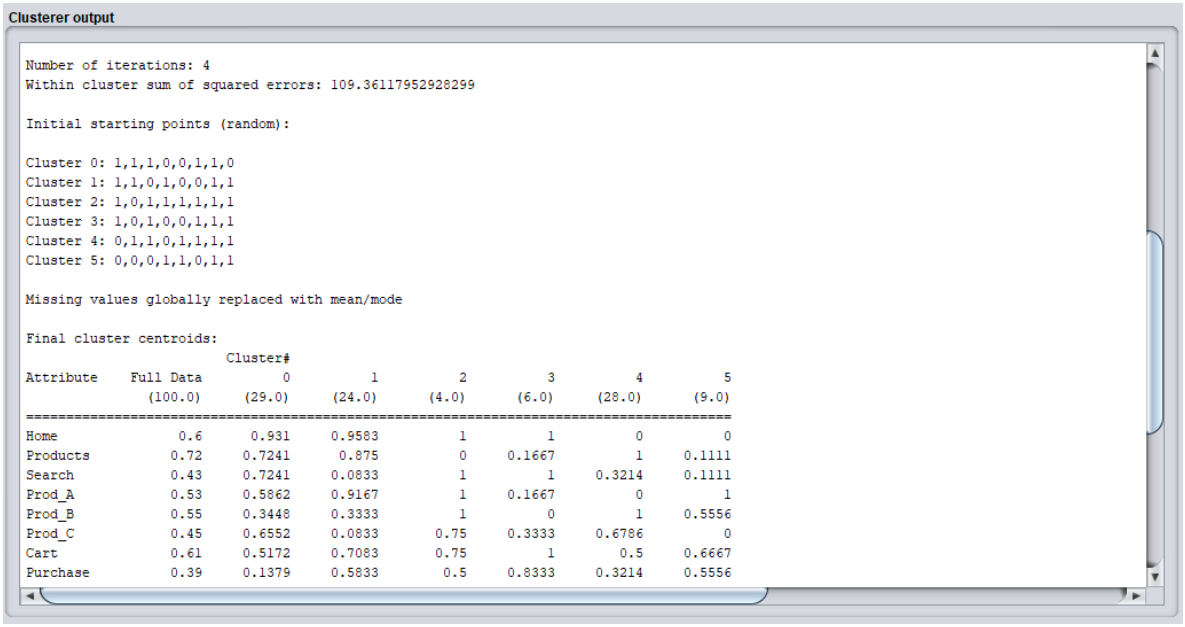
Hình 2: $k = 4$.



Clusterer output

```
Number of iterations: 8
Within cluster sum of squared errors: 113.58260073260074

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,0,1,1
Cluster 4: 0,1,1,0,1,1,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4
               (100.0)     (26.0)     (27.0)      (5.0)     (14.0)     (28.0)
=================================================================================
Home                0.6     0.9615     0.6667          1     0.8571          0
Products           0.72     0.6923     0.6667          0     0.5714          1
Search             0.43     0.6538          0          1     0.8571     0.3214
Prod_A             0.53     0.4615      0.963          1     0.7143          0
Prod_B             0.55     0.3846     0.4444        0.8     0.0714          1
Prod_C             0.45     0.5385          0        0.8     0.5714     0.6786
Cart               0.61     0.4615     0.6296        0.8          1        0.5
Purchase           0.39          0     0.5185        0.4          1     0.3214
```

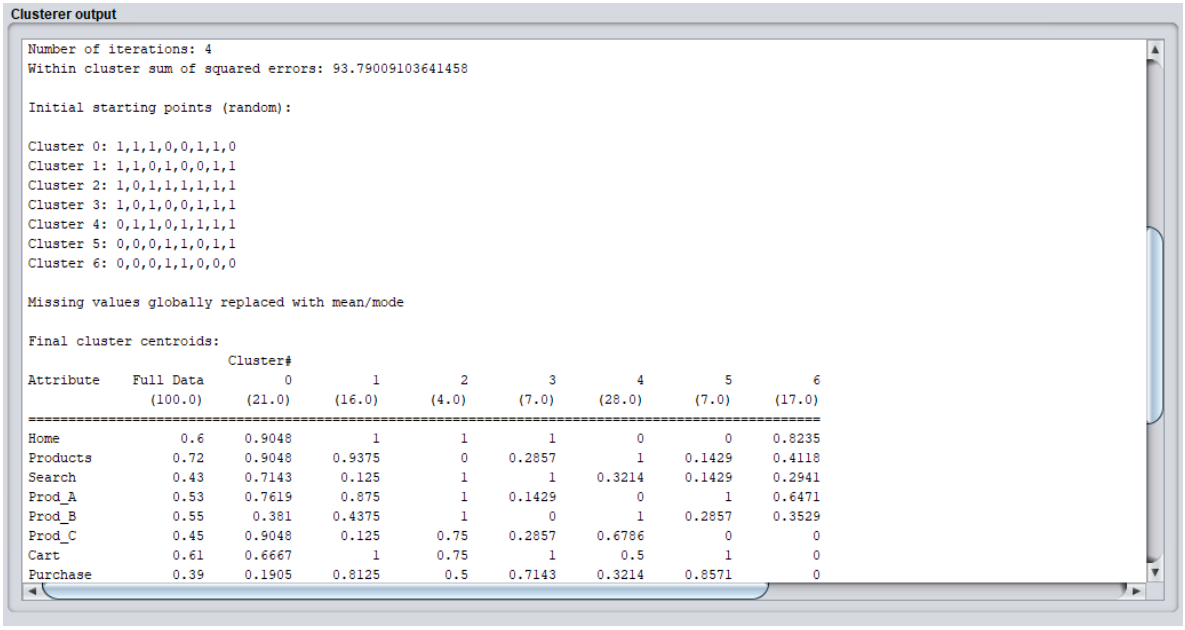
Hình 3: $k = 5$.



Clusterer output

```
Number of iterations: 4
Within cluster sum of squared errors: 109.36117952928299

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1
Cluster 4: 0,1,1,0,1,1,1,1
Cluster 5: 0,0,0,1,1,0,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5
               (100.0)     (29.0)     (24.0)      (4.0)      (6.0)     (28.0)      (9.0)
=================================================================================================
Home                0.6      0.931     0.9583          1          1          0          0
Products           0.72     0.7241      0.875          0     0.1667          1     0.1111
Search             0.43     0.7241     0.0833          1          1     0.3214     0.1111
Prod_A             0.53     0.5862     0.9167          1     0.1667          0          1
Prod_B             0.55     0.3448     0.3333          1          0          1     0.5556
Prod_C             0.45     0.6552     0.0833       0.75     0.3333     0.6786          0
Cart               0.61     0.5172     0.7083       0.75          1        0.5     0.6667
Purchase           0.39     0.1379     0.5833        0.5     0.8333     0.3214     0.5556
```

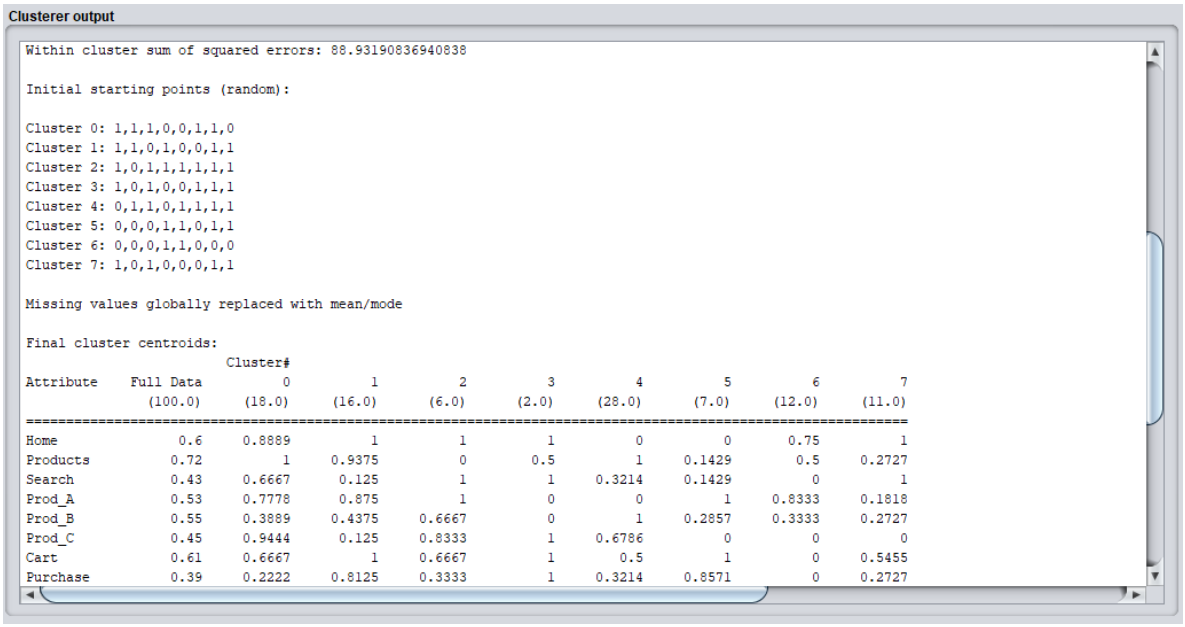
Hình 4: $k = 6$.



Clusterer output

```
Number of iterations: 4
Within cluster sum of squared errors: 93.79009103641458

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1
Cluster 4: 0,1,1,0,1,1,1,1
Cluster 5: 0,0,0,1,1,0,1,1
Cluster 6: 0,0,0,1,1,0,0,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5          6
                (100.0)     (21.0)     (16.0)      (4.0)      (7.0)     (28.0)      (7.0)     (17.0)
=============================================================================================================
Home                 0.6     0.9048          1          1          1          0          0     0.8235
Products            0.72     0.9048     0.9375          0     0.2857          1     0.1429     0.4118
Search              0.43     0.7143      0.125          1          1     0.3214     0.1429     0.2941
Prod_A              0.53     0.7619      0.875          1     0.1429          0          1     0.6471
Prod_B              0.55      0.381     0.4375          1          0          1     0.2857     0.3529
Prod_C              0.45     0.9048      0.125       0.75     0.2857     0.6786          0          0
Cart                0.61     0.6667          1       0.75          1        0.5          1          0
Purchase            0.39     0.1905     0.8125        0.5     0.7143     0.3214     0.8571          0
```

Hình 5: $k = 7$.



Clusterer output

```
Within cluster sum of squared errors: 88.93190836940838

Initial starting points (random):

Cluster 0: 1,1,1,0,0,1,1,0
Cluster 1: 1,1,0,1,0,0,1,1
Cluster 2: 1,0,1,1,1,1,1,1
Cluster 3: 1,0,1,0,0,1,1,1
Cluster 4: 0,1,1,0,1,1,1,1
Cluster 5: 0,0,0,1,1,0,1,1
Cluster 6: 0,0,0,1,1,0,0,0
Cluster 7: 1,0,1,0,0,0,1,1

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4          5          6          7
                (100.0)     (18.0)     (16.0)      (6.0)      (2.0)     (28.0)      (7.0)     (12.0)     (11.0)
=====================================================================================================================
Home                 0.6     0.8889          1          1          1          0          0       0.75          1
Products            0.72          1     0.9375          0        0.5          1     0.1429        0.5     0.2727
Search              0.43     0.6667      0.125          1          1     0.3214     0.1429          0          1
Prod_A              0.53     0.7778      0.875          1          0          0          1     0.8333     0.1818
Prod_B              0.55     0.3889     0.4375     0.6667          0          1     0.2857     0.3333     0.2727
Prod_C              0.45     0.9444      0.125     0.8333          1     0.6786          0          0          0
Cart                0.61     0.6667          1     0.6667          1        0.5          1          0     0.5455
Purchase            0.39     0.2222     0.8125     0.3333          1     0.3214     0.8571          0     0.2727
```

Hình 6: $k = 8$.

Từ các câu tiếp theo, nhóm chọn $k = 5$ để trả lời.

2.1.2 Câu 2

- New-user: [1, 0, 1, 0, 1, 0, 0, 0]
- Khoảng cách giữa điểm mới với các cụm: {1.3, 1.859, 1.574, 2.0, 1.809}

→ New-user như mô tả của câu này thuộc cụm 1 (cluster0). Ta có $Prod\_A = 0.4615 < Prod\_C = 0.5385$ nên sẽ giới thiệu $Prod\_C$ đến người dùng này.

2.1.3 Câu 3

- New-user: [0, 1, 0, 0, 0, 1, 0, 0]
- Khoảng cách giữa điểm mới với các cụm: {1.49, 1.829, 2.3409, 2.08, 1.24}

→ New-user như mô tả của câu này thuộc cụm 5 (cluster4). Ta có $Prod\_B = 1 < Prod\_A = 0$ nên sẽ giới thiệu $Prod\_B$ đến người dùng này.

2.1.4 Câu 4

- Người dùng thông thường (window shopper, xem nhiều sản phẩm): Cụm 3 (cluster2) vì tỉ lệ xem $Prod\_A$, $Prod\_B$ và $Prod\_C$ cao hơn hẳn các cụm còn lại.
  Dẫn chứng:
  + Mẫu 09: [1,0,1,1,1,1,1,1]
  + Mẫu 10: [1,0,1,1,1,1,1,0]
  + Mẫu 14: [1,0,1,1,0,1,1,0]
- Người dùng tập trung (biết rõ cần mua sản phẩm gì): Cụm 2 (cluster1) vì có tỉ lệ $Cart$ và $Purchase$ cao trong khi $Search$ thấp.
  Dẫn chứng:
  + Mẫu 69: [1,1,0,1,0,0,1,1]
  + Mẫu 71: [1,1,0,1,1,0,1,1]
  + Mẫu 72: [0,0,0,1,0,0,1,1]
- Người dùng tìm kiếm (sử dụng chức năng search để tìm sản phẩm cần mua): Cụm 3 (cluster2) vì có tỉ lệ người dùng sử dụng chức năng search là 100%.
  Dẫn chứng:
  + Mẫu 09: [1,0,1,1,1,1,1,1]
  + Mẫu 10: [1,0,1,1,1,1,1,0]
  + Mẫu 14: [1,0,1,1,1,1,0,0]

2.1.5 Câu 5

- Sản phẩm đơn lẻ - cụm 2 (cluster1): $Prod\_A$.
  + Đặc điểm hành vi duyệt trang: vào thẳng trang $Prod\_A$.

+ Xu hướng thanh toán: trung bình (khoảng 51%).
+ Dẫn chứng:
  * Mẫu 53: [1,1,0,1,0,0,0,0]
  * Mẫu 54: [0,1,0,1,1,0,0,0]
  * Mẫu 55: [0,0,0,1,1,0,1,1]

– Sản phẩm đơn lẻ - cụm 5 (cluster4): $Prod\_B$.
+ Đặc điểm hành vi duyệt trang: $Product \rightarrow Prod\_B$.
+ Xu hướng thanh toán: trung bình (khoảng 32%).
+ Dẫn chứng:
  * Mẫu 82: [0,1,1,0,1,0,0,0]
  * Mẫu 83: [0,1,0,0,1,1,1,1]
  * Mẫu 84: [0,1,0,0,1,1,0,0]

– Nhóm các sản phẩm - cụm 3 (cluster2): $Prod\_A$, $Prod\_B$, $Prod\_C$.
+ Đặc điểm hành vi duyệt trang: $Home \rightarrow Search \rightarrow$ các trang $Prod\_A$, $Prod\_B$, $Prod\_C$.
+ Xu hướng thanh toán: trung bình (khoảng 40%).
+ Dẫn chứng:
  * Mẫu 09: [1,0,1,1,1,1,1,1]
  * Mẫu 10: [1,0,1,1,1,1,1,0]
  * Mẫu 14: [1,0,1,1,1,1,0,0]

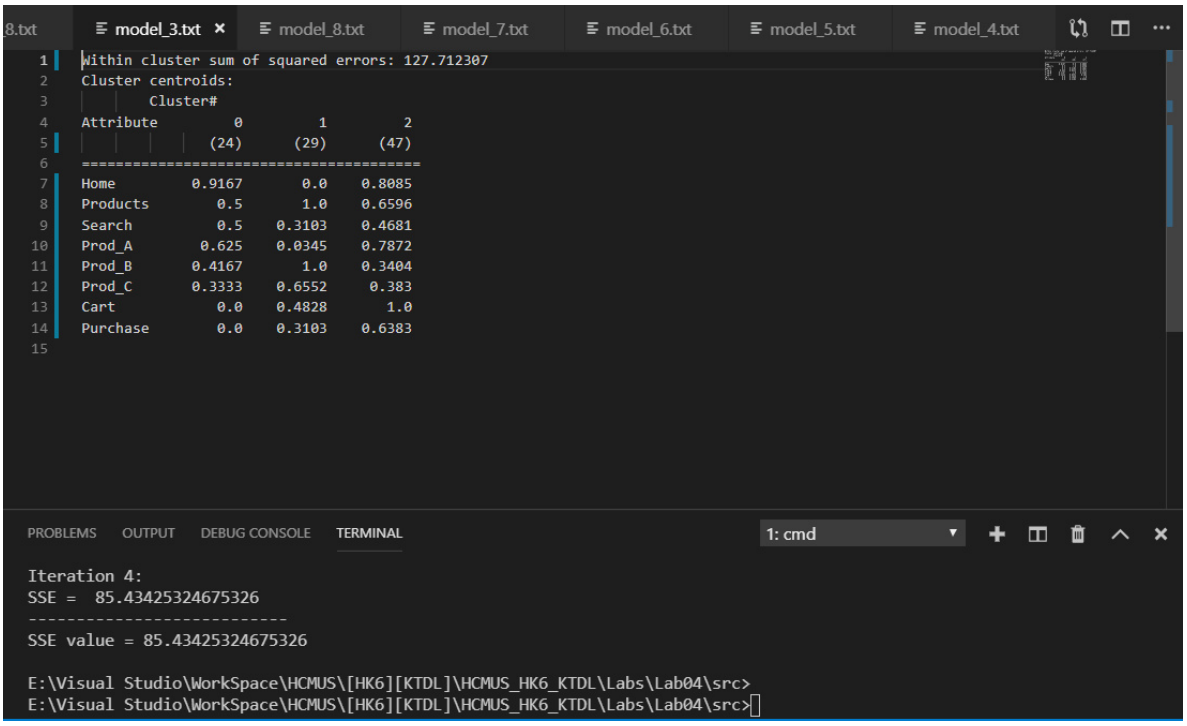
2.1.6 Câu 6

Có thể, đó là cụm 5 (cluster4) vì có tỉ lệ vào 3 trang $Home$, $Product$, $Search$ thấp. Chiến dịch quảng cáo cho$Prod_B$ thành công hơn vì có tỉ lệ truy cập ở cụm này là 100%.

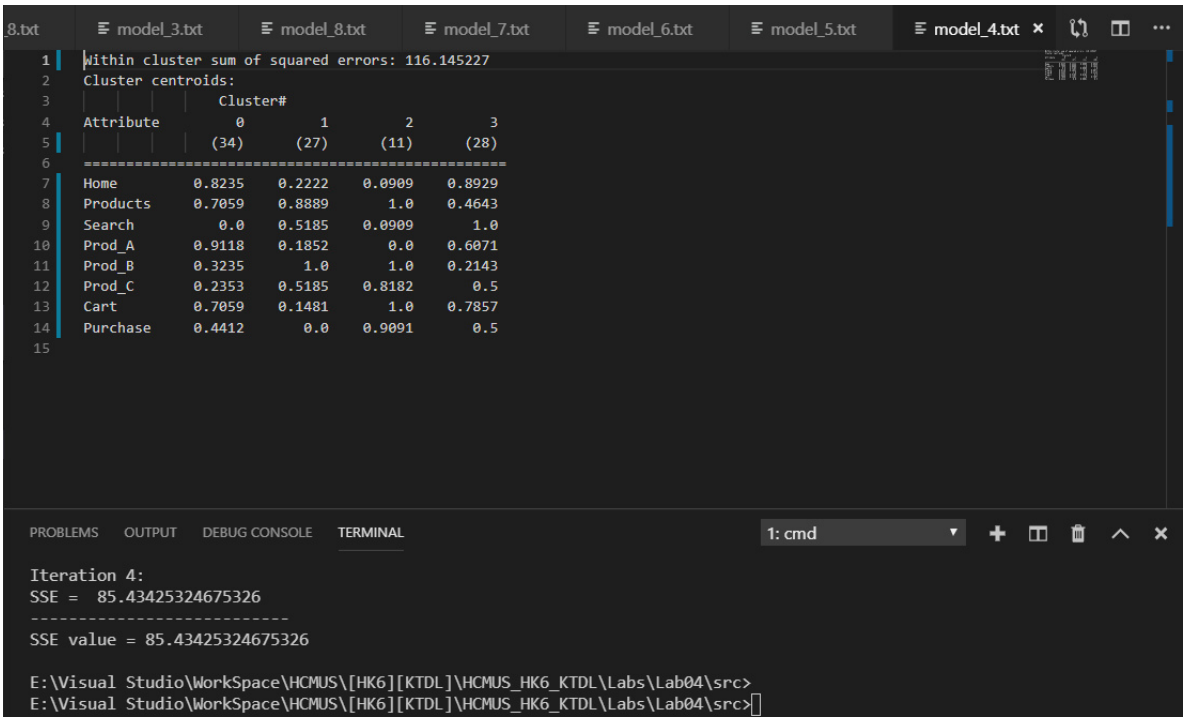
2.2 Phần cài đặt

Chạy chương trình cài đặt với tập dữ liệu `sessions.csv`. Đối chiếu kết quả phát sinh được với kết quả của Weka trên cùng giá trị k (từ 3 đến 8).

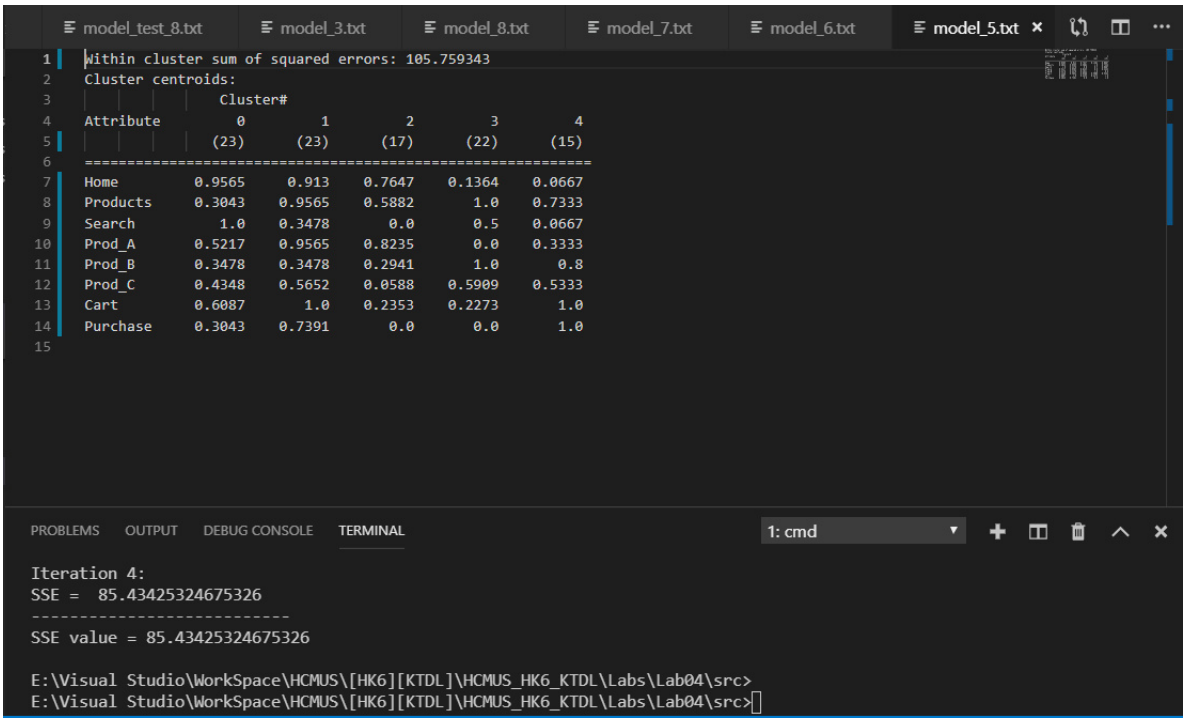
2.2.1 Kết quả chạy với *random-initial-starting-points*



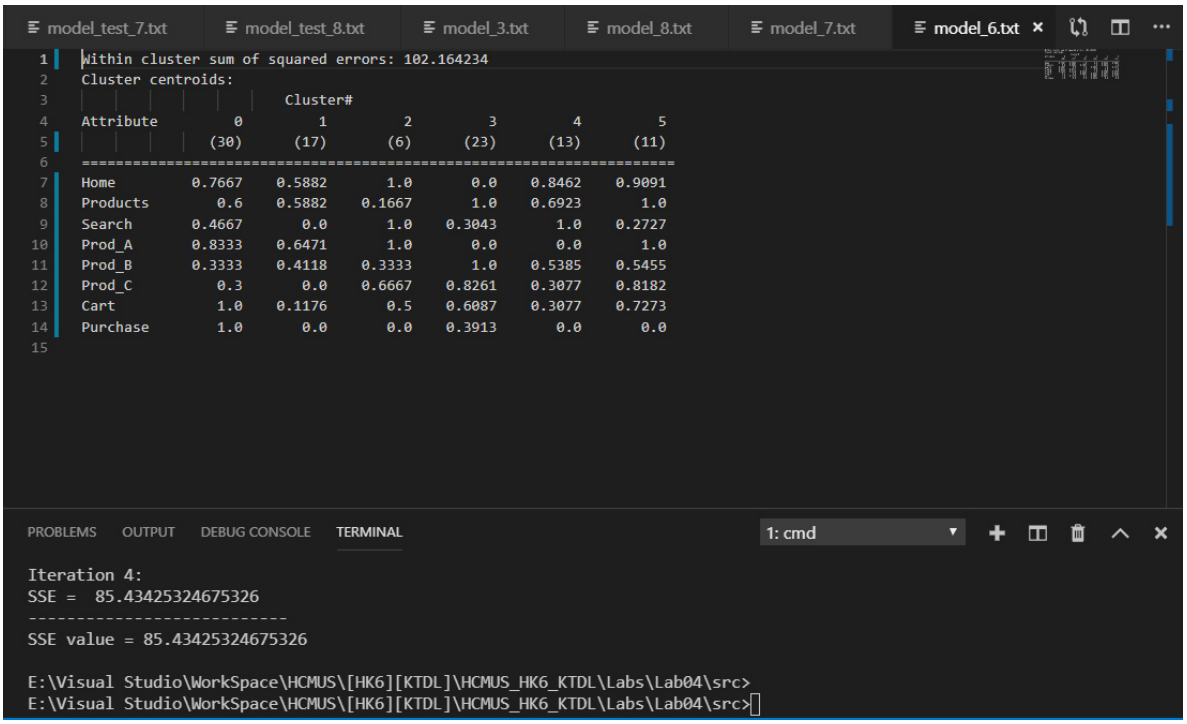
Hình 7: Random-initial-starting-points: $k = 3$.



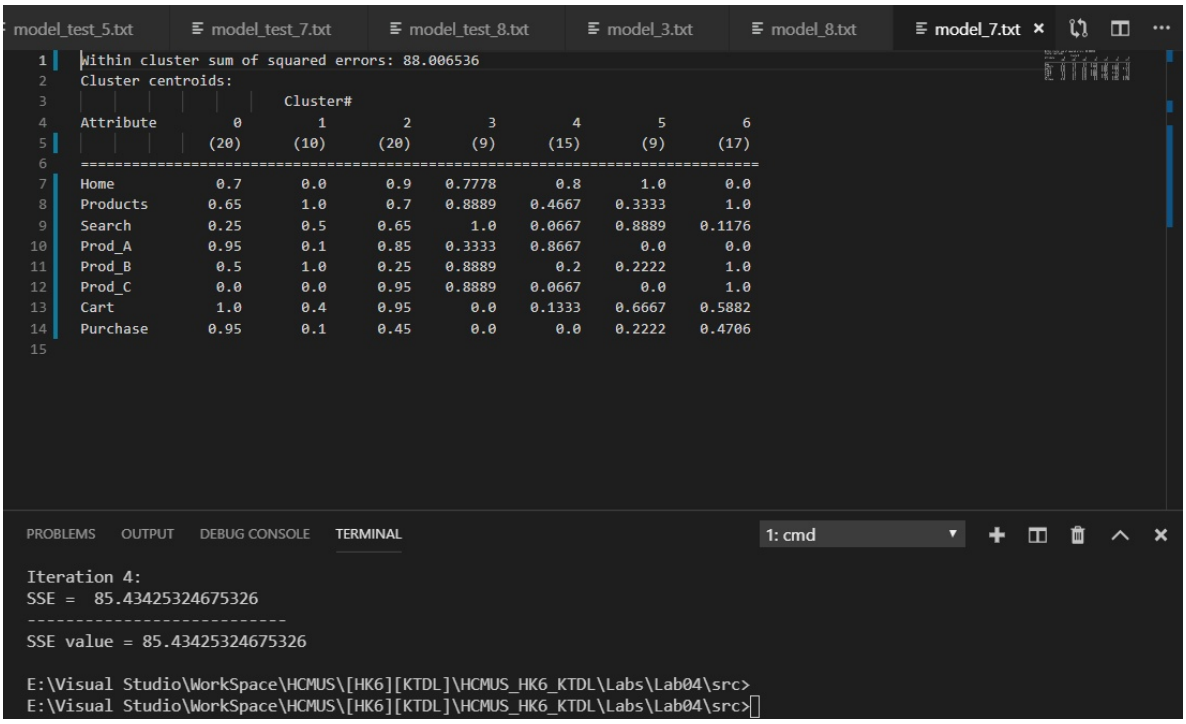
Hình 8: Random-initial-starting-points: $k = 4$.



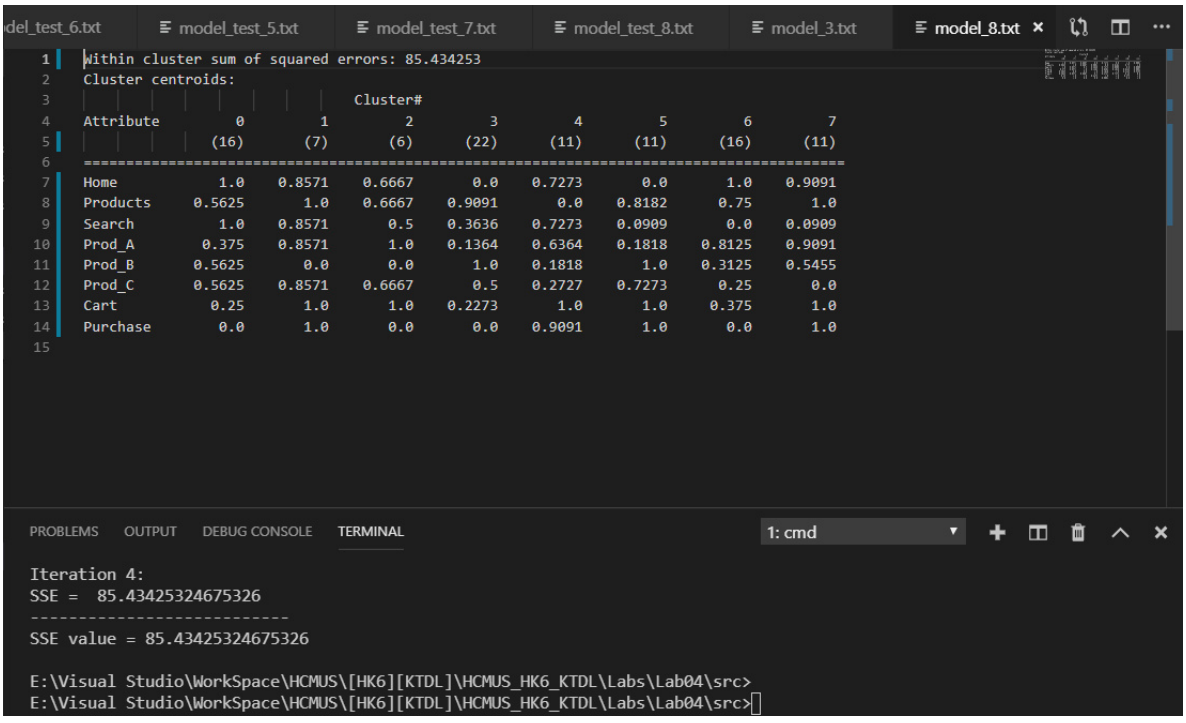
Hình 9: Random-initial-starting-points: $k = 5$.



Hình 10: Random-initial-starting-points: $k = 6$.

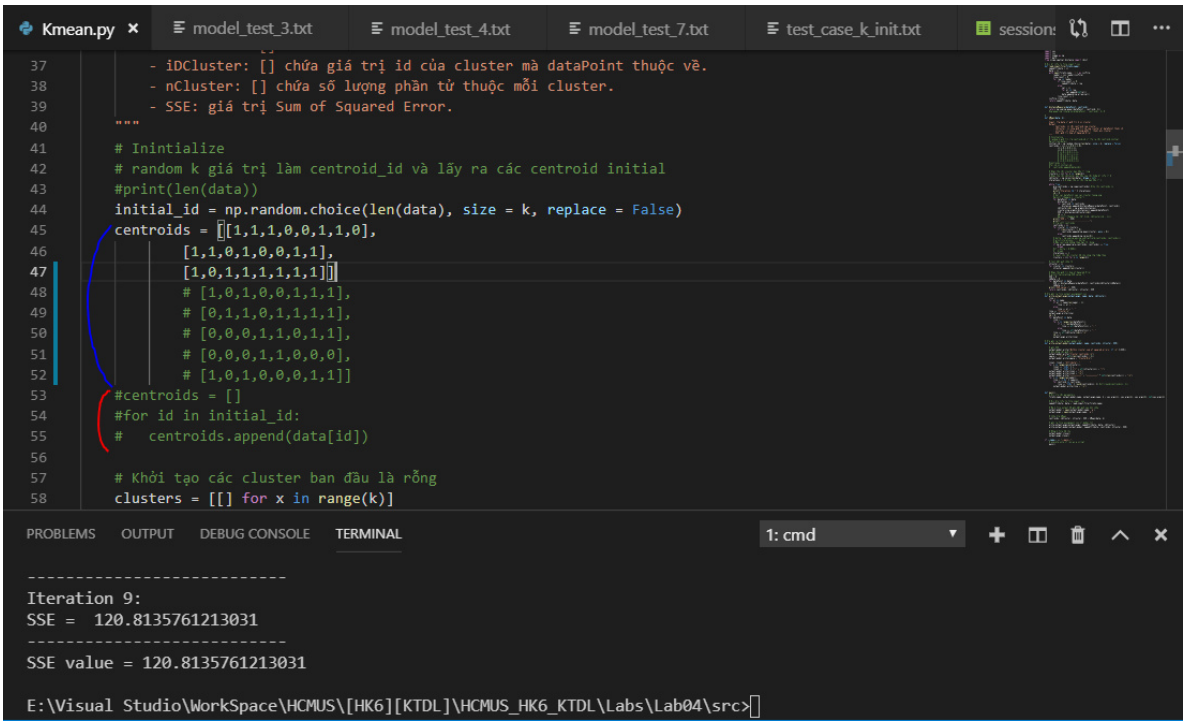


Hình 11: Random-initial-starting-points: $k = 7$.

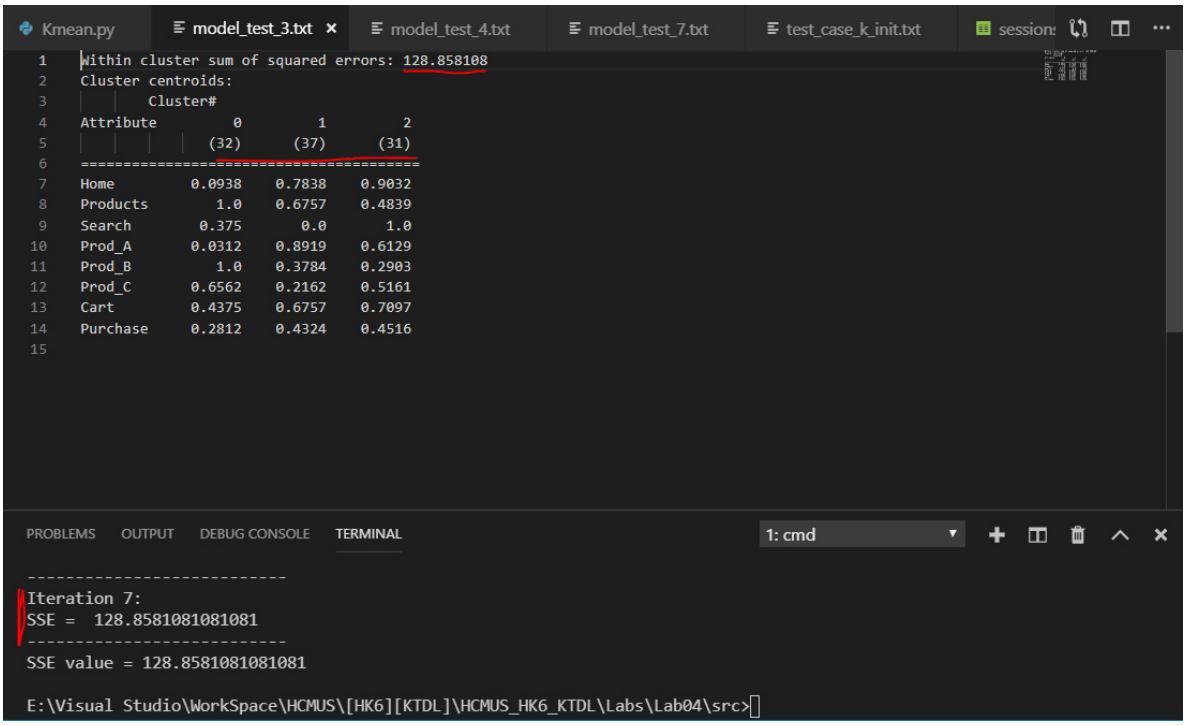


Hình 12: Random-initial-starting-points: $k = 8$.

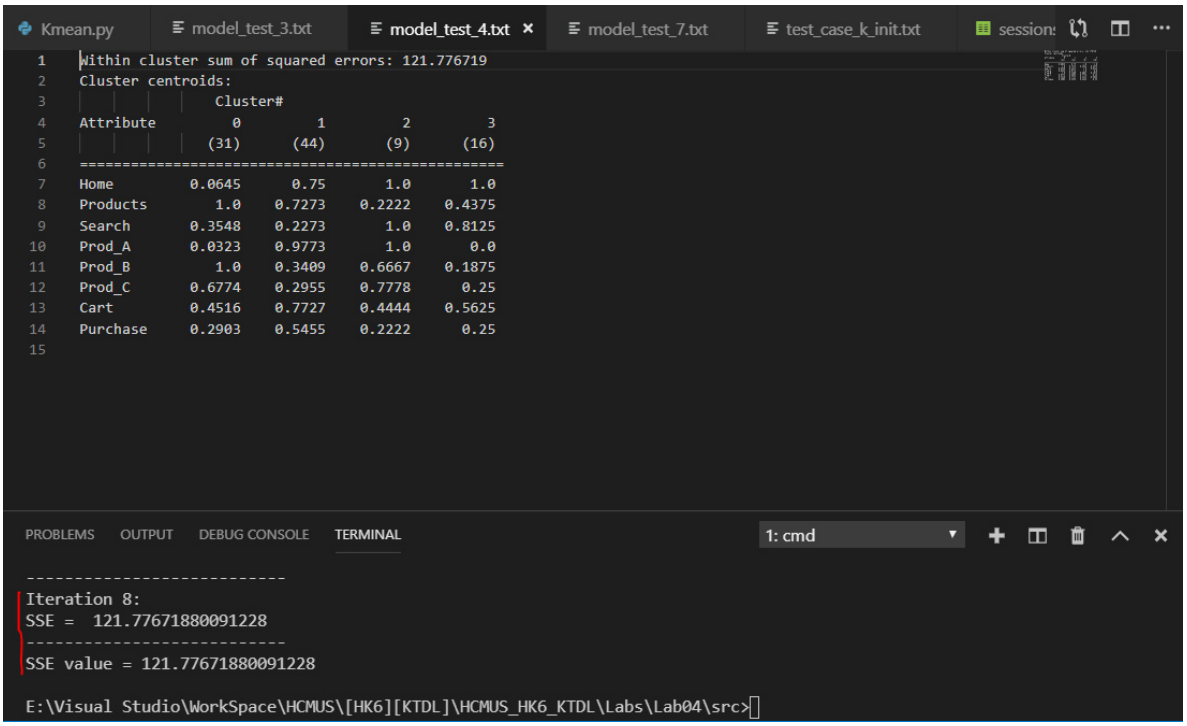
2.2.2 Kết quả chạy với *Weka-initial-starting-points*



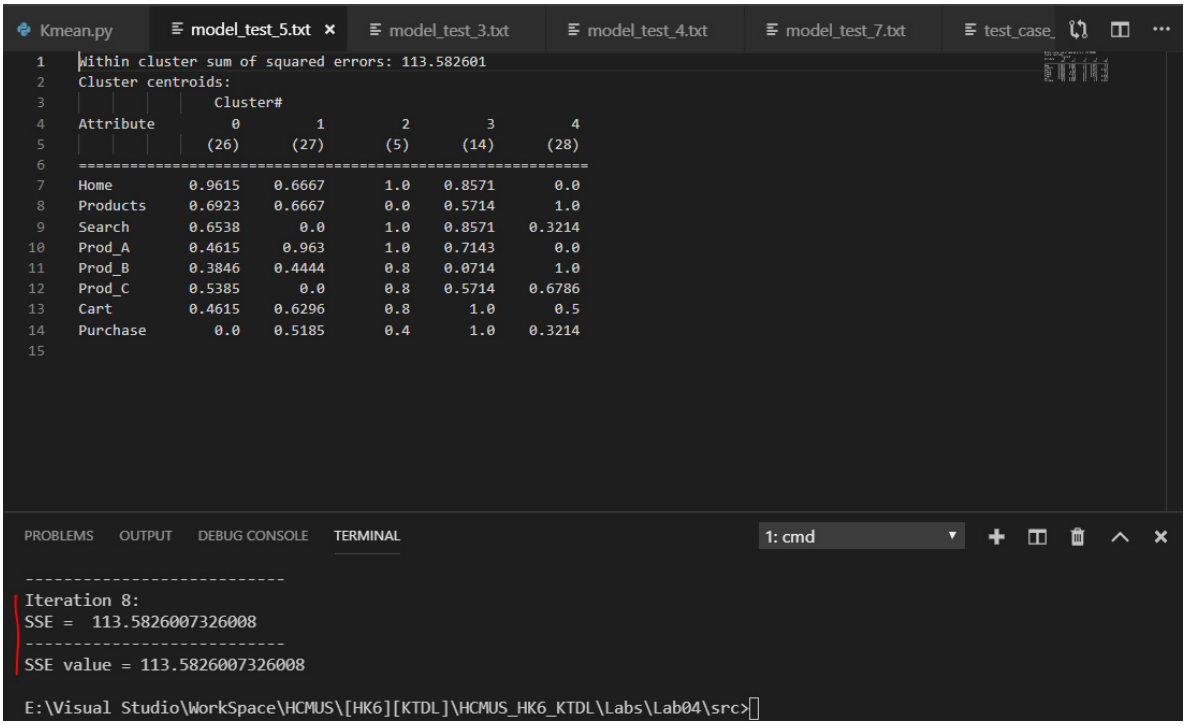
Hình 13: Ở đoạn code này, nhóm thiết lập các initial-starting-points giống với bên Weka để kiểm tra phần cài đặt giải thuật k-mean của nhóm (comment phần màu đỏ, copy từ file test_case_k_init.txt ứng với số lượng k cụm muốn phân).



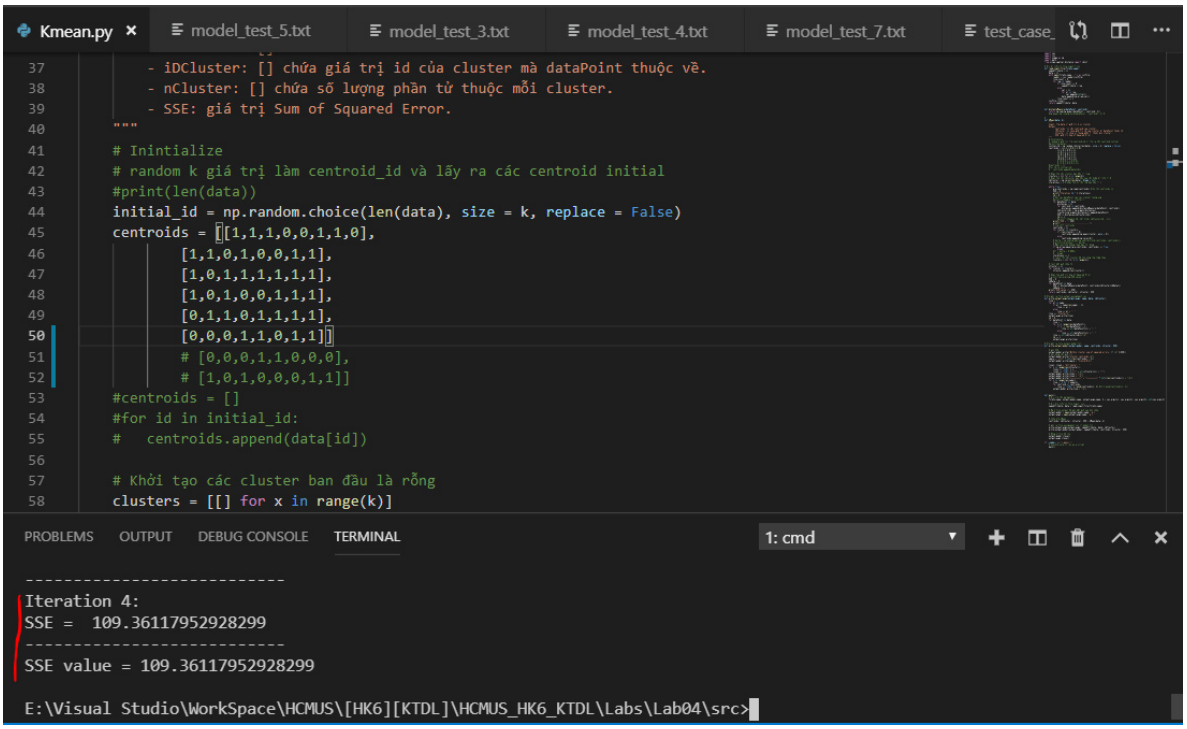
Hình 14: Weka-initial-starting-points: $k = 3$.



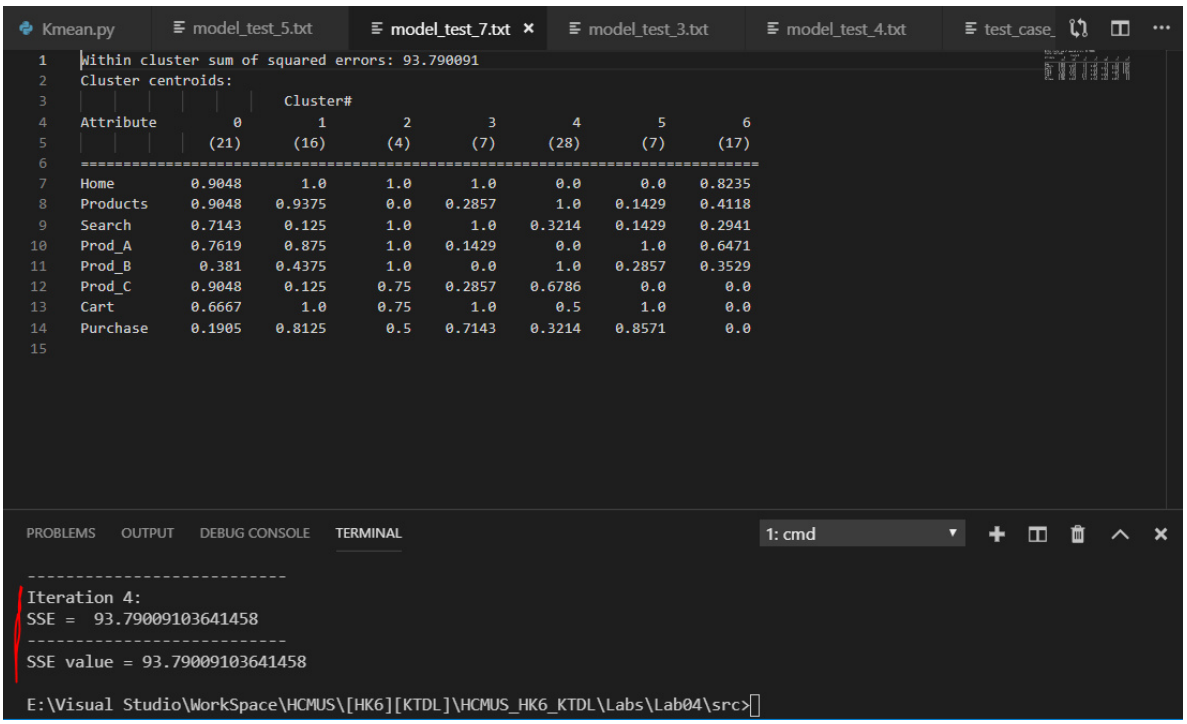
Hình 15: Weka-initial-starting-points: $k = 4$.



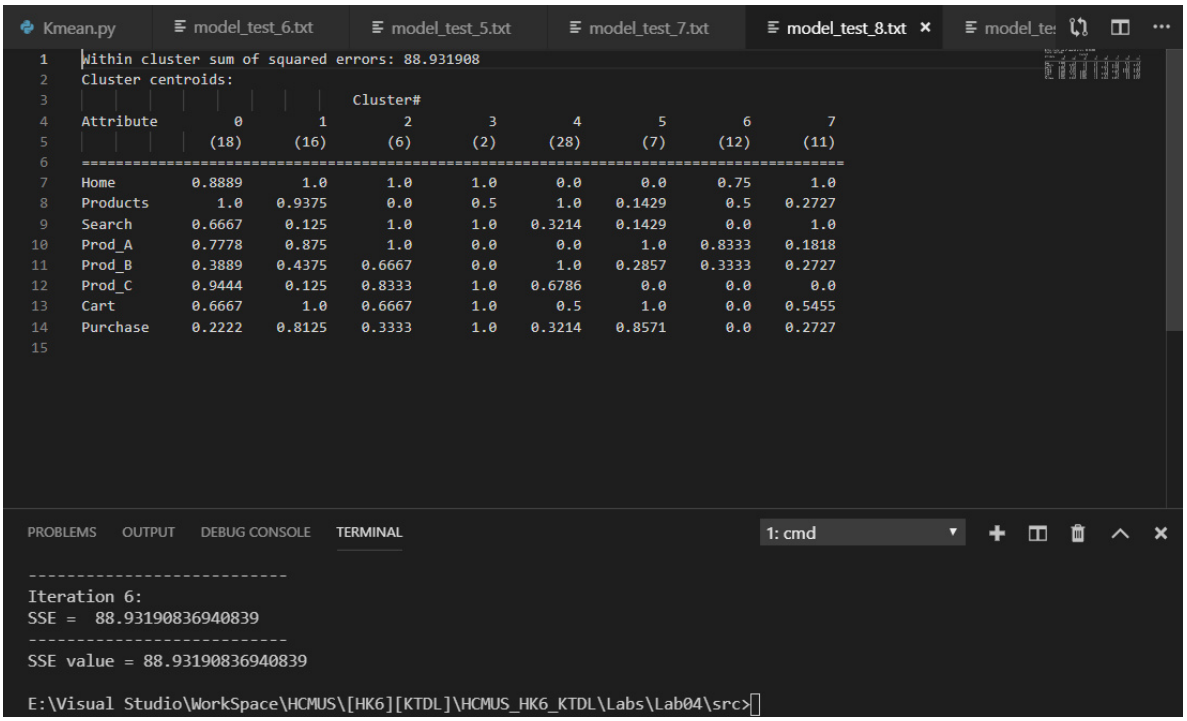
Hình 16: Weka-initial-starting-points: $k = 5$.



Hình 17: Weka-initial-starting-points: $k = 6$.



Hình 18: Weka-initial-starting-points: $k = 7$.



Hình 19: Weka-initial-starting-points: $k = 8$.

2.2.3 Nhận xét

- Kết quả gom cụm và SSE giữa phần cài đặt của nhóm và Weka khác nhau vì các điểm khởi tạo ban đầu (init centroid) của nhóm được lấy một cách ngẫu nhiên từ tập dữ liệu, việc lấy random khác so với Weka dẫn đến lấy các điểm khởi tạo ban đầu khác. Trong thực tế, tùy vào các center ban đầu mà thuật toán có thể có tốc độ hội tụ rất chậm hoặc thậm chí cho chúng ta nghiệm không chính xác (chỉ là local minimum - điểm cực tiểu - mà không phải giá trị nhỏ nhất). Do đó người ta thường chạy k-mean nhiều lần với các center ban đầu khác nhau rồi chọn cách có hàm mất mát cuối cùng đạt giá trị nhỏ nhất.
- Khi nhóm thiết lập các initial starting points giống với bên Weka thì kết quả gom cụm và SSE giống với kết quả Weka thực hiện.

3 Đánh giá

STT	Nội dung	Hoàn thành
1	Phần Weka	100%
2	Phần cài đặt	100%
Mức độ hoàn thành tổng thể của bài tập:		**100%**

Tài liệu

[1] Slide lý thuyết.

[2] Trang chủ của Weka: http://www.cs.waikato.ac.nz/ml/weka/

[3] J. Han and M. Kamber, *Data Mining, Concepts and Techniques, Second Edition*