TUGAS 7

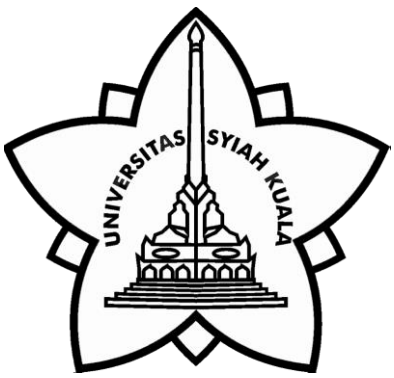
Disusun Untuk Memenuhi

Tugas Data Mining

Oleh:

Khairul Umam Albi

2008107010072



JURUSAN INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA
DARUSSALAM, BANDA ACEH
2022

MELAKUKAN PROSES TERHADAP DATASET Seeds dan Iris MENGGUNAKAN METODE K-Means.

1. Tools yang digunakan
    - Aplikasi Weka
    - Text editor(notepad dan vscode)
    - Kalkulator
2. Langkah Proses Klasifikasi dataset Seeds
    - Dataset dapat diunduh melalui link berikut : http://archive.ics.uci.edu/ml/datasets/seeds
    - Menganti dan mengubah file menjadi bentuk arrf
    - Buka applikasi weka dan pilih opsi open didalam aplikasi tersebut.
    - Pilih tab cluster dan pilih metode SimpleKMeans
    - Pilih opsi test option dengan K = 2 dan K = 3 dan lakukan dengan fungsi perhitungan jarak Manhattan dan Euclidian Distance
    - tekan start.
3. Langkah Proses Klasifikasi dataset iris
    - Dataset dapat diambil dari folder weka : D:\Program Files\Weka-2-8-6\data –jika menggunakan OS Windows
    - Buka applikasi weka dan pilih opsi open didalam aplikasi tersebut.
    - Pilih tab cluster dan pilih metode SimpleKMeans
    - Pilih opsi test option dengan K = 2 dan K = 3 dan lakukan dengan fungsi perhitungan jarak Manhattan dan Euclidian Distance
    - tekan start.

A. Dataset seeds menggunakan k = 2 dan k = 3 dengan perhitungan Manhatan

- K = 2

```
kMeans
======

Number of iterations: 6
Sum of within cluster distances: 181.61203521676867

Initial starting points (random):

Cluster 0: 18.59,16.05,0.9066,6.037,3.86,6.001,5.877
Cluster 1: 10.93,12.8,0.839,5.046,2.717,5.398,5.045

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                 Full Data           0           1
                            (210.0)      (76.0)     (134.0)
============================================================
area_A                      14.355       18.57      12.725
perimeter_P                  14.32      16.185       13.58
compactness_C               0.8735      0.8826      0.8658
length_of_kernel            5.5235       6.126      5.3405
width_of_kernel              3.237      3.6855       3.026
asymmetry_coefficient        3.599      3.4225      3.6345
length_of_kernel_groove      5.223      5.9655       5.089



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

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

Clustered Instances

0       76 ( 36%)
1      134 ( 64%)
```

Hasil clustering menggunakan simple k-means menghasilkan 2 kelompok kluster yang mana kelompok 0 sebanyak 36 % dan kelompok 1 sebanyak 64 %. Nomor iterasi berhenti di iterasi ke-6 dan cluster distances sebanyak : 181,61203521676867

- K = 3

```
Number of iterations: 8
Sum of within cluster distances: 139.87703488668768

Initial starting points (random):

Cluster 0: 18.59,16.05,0.9066,6.037,3.86,6.001,5.877
Cluster 1: 10.93,12.8,0.839,5.046,2.717,5.398,5.045
Cluster 2: 13.32,13.94,0.8613,5.541,3.073,7.035,5.44

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                 Full Data           0           1           2
                            (210.0)     (63.0)      (79.0)      (68.0)
=============================================================================
area_A                      14.355      18.81       12.01       14.695
perimeter_P                  14.32      16.26       13.27       14.505
compactness_C               0.8735      0.885      0.8511        0.882
length_of_kernel            5.5235      6.172       5.226      5.5745
width_of_kernel              3.237      3.755       2.847      3.2895
asymmetry_coefficient        3.599      3.477        4.67      2.6995
length_of_kernel_groove      5.223      6.053       5.088      5.1625



Time taken to build model (full training data) : 0.02 seconds

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

Clustered Instances

0       63 ( 30%)
1       79 ( 38%)
2       68 ( 32%)
```

Hasil clustering menggunakan simple k-means menghasilkan 3 kelompok kluster yang mana kelompok 0 sebanyak 30% ,kelompok 1 sebanyak 38%, dan kelompok 2 sebanyak 32%. Nomor iterasi berhenti di iterasi ke-8 dan cluster distances sebanyak : 139.87703488668768

B. Dataset seeds menggunakan k = 2 dan k = 3 dengan perhitungan Euclidian Distance

- K = 2

```
Number of iterations: 8
Within cluster sum of squared errors: 34.81326792694563

Initial starting points (random):

Cluster 0: 18.59,16.05,0.9066,6.037,3.86,6.001,5.877
Cluster 1: 10.93,12.8,0.839,5.046,2.717,5.398,5.045

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                 Full Data           0           1
                            (210.0)     (77.0)     (133.0)
=============================================================
area_A                      14.8475    18.1586    12.9306
perimeter_P                 14.5593    16.0548    13.6935
compactness_C                 0.871     0.8838     0.8636
length_of_kernel             5.6285     6.1274     5.3397
width_of_kernel              3.2586     3.6605     3.0259
asymmetry_coefficient        3.7002     3.4804     3.8274
length_of_kernel_groove      5.4081     5.9717     5.0817



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

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

Clustered Instances

0       77 ( 37%)
1      133 ( 63%)
```

Hasil clustering menggunakan simple k-means menghasilkan 2 kelompok kluster yang mana kelompok 0 sebanyak 37 % dan kelompok 1 sebanyak 63 %. Nomor iterasi berhenti di iterasi ke-8 dan cluster sum of squared errors: 34.81326792694563

- K = 3

```
======

Number of iterations: 5
Within cluster sum of squared errors: 22.024363075666038

Initial starting points (random):

Cluster 0: 18.59,16.05,0.9066,6.037,3.86,6.001,5.877
Cluster 1: 10.93,12.8,0.839,5.046,2.717,5.398,5.045
Cluster 2: 13.32,13.94,0.8613,5.541,3.073,7.035,5.44

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                 Full Data           0           1           2
                            (210.0)      (64.0)      (77.0)      (69.0)
=================================================================================
area_A                     14.8475     18.6102     11.8961     14.6512
perimeter_P                14.5593     16.2517     13.2577      14.442
compactness_C                0.871      0.8846      0.8498      0.8821
length_of_kernel            5.6285      6.1955      5.2306      5.5467
width_of_kernel             3.2586      3.7096       2.858      3.2873
asymmetry_coefficient       3.7002      3.5921      4.5995      2.7969
length_of_kernel_groove     5.4081      6.0567      5.0862      5.1656



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

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

Clustered Instances

0       64 ( 30%)
1       77 ( 37%)
2       69 ( 33%)
```

Hasil clustering menggunakan simple k-means menghasilkan 3 kelompok kluster yang mana kelompok 0 sebanyak 37 % , kelompok 1 sebanyak 63 %, dan kelompok 2 sebanyak 69% . Nomor iterasi berhenti di iterasi ke-5 dan cluster sum of squared errors: 22.024363075666038

C. Dataset iris menggunakan k = 2 dan k = 3 dengan perhitungan Manhatan

- K = 2

```
kMeans
======

Number of iterations: 6
Sum of within cluster distances: 63.72622410546139

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1
                 (150.0)     (98.0)     (52.0)
================================================
sepallength          5.8        6.3          5
sepalwidth             3        2.9        3.4
petallength         4.35        4.9        1.5
petalwidth           1.3        1.6        0.2



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

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

Clustered Instances

0       98 ( 65%)
1       52 ( 35%)
```

Hasil clustering menggunakan simple k-means menghasilkan 2 kelompok kluster yang mana kelompok 0 sebanyak 65 % dan kelompok 1 sebanyak 52 %. Nomor iterasi berhenti di iterasi ke-6 dan cluster distances sebanyak : 63.72622410546139

- K = 3

```
kMeans
======

Number of iterations: 5
Sum of within cluster distances: 47.779425612052705

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3
Cluster 2: 6.9,3.1,5.1,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2
                 (150.0)     (62.0)     (50.0)     (38.0)
=========================================================
sepallength          5.8        5.9          5        6.7
sepalwidth             3        2.8        3.4          3
petallength         4.35        4.5        1.5       5.65
petalwidth           1.3        1.4        0.2        2.1



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

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

Clustered Instances

0       62 ( 41%)
1       50 ( 33%)
2       38 ( 25%)
```

Hasil clustering menggunakan simple k-means menghasilkan 2 kelompok kluster yang mana kelompok 0 sebanyak 62 % dan kelompok 1 sebanyak 50 %, dan kelompok 2 sebanyak 38%. Nomor iterasi berhenti di iterasi ke-5 dan cluster distances sebanyak : 47.779425612052705

D. Dataset iris menggunakan k = 2 dan k = 3 dengan perhitungan Euclidian Distance

- K = 2

```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 12.143688281579722

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data           0           1
                 (150.0)     (100.0)      (50.0)
================================================
sepallength       5.8433       6.262       5.006
sepalwidth         3.054       2.872       3.418
petallength       3.7587       4.906       1.464
petalwidth        1.1987       1.676       0.244



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

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

Clustered Instances

0      100 ( 67%)
1       50 ( 33%)
```

Hasil clustering menggunakan simple k-means menghasilkan 2 kelompok kluster yang mana kelompok 0 sebanyak 67 % dan kelompok 1 sebanyak 33 %. Nomor iterasi berhenti di iterasi ke-7 dan cluster sum of squared errors: 12.143688281579722.

- K = 3

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 6.998114004826762

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4
Cluster 1: 6.2,2.9,4.3,1.3
Cluster 2: 6.9,3.1,5.1,2.3

Missing values globally replaced with mean/mode

Final cluster centroids:
                           Cluster#
Attribute      Full Data          0          1          2
                 (150.0)     (61.0)     (50.0)     (39.0)
============================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795



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

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

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

Hasil clustering menggunakan simple k-means menghasilkan 2 kelompok kluster yang mana kelompok 0 sebanyak 41 % dan kelompok 1 sebanyak 50 %, dan kelompok 2 sebanyak 26 %. Nomor iterasi berhenti di iterasi ke-6 dan cluster sum of squared errors: 6.998114004826762.

E. Rangkuman

Dari hasil klustering tersebut didapatkan bahwa iterasi, distance,dan cluster sum of squared errors yang digunakan dengan perhitungan jarak Manhattan dan Euclidian Distance terdapat kesamaan(similiarty) dan ketidaksamaan (dissimilarity) dalam dataset iris dan seeds yang menggunakan k=2 dan k=3.