TUGAS BESAR TAHAP 1 CLUSTERING

Laporan

Dibuat untuk memenuhi tugas mata kuliah Pembelajaran Mesin

oleh :

Farhan Anas (1301183427)

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS INFORMATIKA

UNIVERSITAS TELKOM

BANDUNG

2021

1. Formulasi Masalah

   Permasalahan yang ingin diselesaikan pada tugas besar tahap 1 ini adalah bagaimana cara mengelompokkan data yang sudah disediakan menjadi beberapa kelompok(Clustering).

2. Eksplorasi dan Persiapan Data

   Terdapat beberapa Teknik pra-pemroresan data yang dilakukan sebelum dilakukan clustering pada dataset, diantaranya yakni :

   a. Remove Duplicate

      Data yang memiliki duplikat didalam dataset perlu dihapus agar mempercepat proses clustering dikarenakan jika terdapat data yang duplikat, hanya 1 data yang memiliki value.

```
dups = df.duplicated()
print('Number of duplicate rows = %d' % (dups.sum()))
```

```
Number of duplicate rows = 585
```

```
df = df.drop_duplicates()
df.describe()
```

	SuhuMin	SuhuMax	Penguapan	Suhu9am	Suhu3pm
count	107949.000000	108114.000000	61745.000000	107732.000000	106371.000000
mean	12.196376	23.215019	5.465474	16.991518	21.673191
std	6.389445	7.106727	4.208063	6.477803	6.923169
min	-8.500000	-4.800000	0.000000	-7.200000	-5.400000
25%	7.600000	17.900000	2.600000	12.300000	16.600000
50%	12.000000	22.600000	4.800000	16.700000	21.100000
75%	16.800000	28.200000	7.400000	21.600000	26.400000
max	33.900000	47.300000	145.000000	40.200000	46.700000

   b. Replace Null

      Data yang memiliki nilai null pada kolomnya didalam dataset perlu diganti dengan sebuah nilai yang dimana dalam hal ini dilakukan terlebih dahulu perhitungan skewness pada tiap kolom yang digunakan dengan syarat jika nilai skewness sebuah kolom diantara -2 sampai 2 maka nilai null direplace dengan menggunakan nilai mean dari kolom tersebut dan jika tidak maka direplace dengan menggunakan nilai median dari kolom tersebut.

```
[48] check_null = df.isnull().sum()
     print(check_null)

     SuhuMin        561
     SuhuMax        396
     Penguapan    46765
     Suhu9am        778
     Suhu3pm       2139
     dtype: int64
```

```
[49] col = ['SuhuMin', 'SuhuMax', 'Penguapan', 'Suhu9am', 'Suhu3pm']
     df[col].skew(axis=0, skipna=True)

     SuhuMin      0.018047
     SuhuMax      0.221010
     Penguapan    3.916798
     Suhu9am      0.085141
     Suhu3pm      0.237995
     dtype: float64
```

```
df['SuhuMin'].fillna(df['SuhuMin'].mean(),inplace=True)
df['SuhuMax'].fillna(df['SuhuMax'].mean(),inplace=True)
df['Penguapan'].fillna(df['Penguapan'].median(),inplace=True)
df['Suhu9am'].fillna(df['Suhu9am'].mean(),inplace=True)
df['Suhu3pm'].fillna(df['Suhu3pm'].mean(),inplace=True)
df.describe()
```

	SuhuMin	SuhuMax	Penguapan	Suhu9am	Suhu3pm
count	108510.000000	108510.000000	108510.000000	108510.000000	108510.000000
mean	12.196376	23.215019	5.178672	16.991518	21.673191
std	6.372906	7.093747	3.191352	6.454539	6.854592
min	-8.500000	-4.800000	0.000000	-7.200000	-5.400000
25%	7.600000	17.900000	4.000000	12.300000	16.700000
50%	12.000000	22.700000	4.800000	16.800000	21.300000
75%	16.800000	28.200000	5.200000	21.500000	26.200000
max	33.900000	47.300000	145.000000	40.200000	46.700000

c. Normalize Data

Tiap data tentunya memiliki berbagai macam range nilai yang dimana kedepannya mungkin saja menimbulkan masalah yakni sebuah kolom yang memiliki range lebih besar akan menjadi lebih dominan daripada kolom yang memiliki range lebih kecil sehingga data di tiap kolom perlu diubah ke dalam nilai dengan range 0 sampai 1.

```
[52] df = (df - df.min()) / (df.max() - df.min())
     df.describe()
```

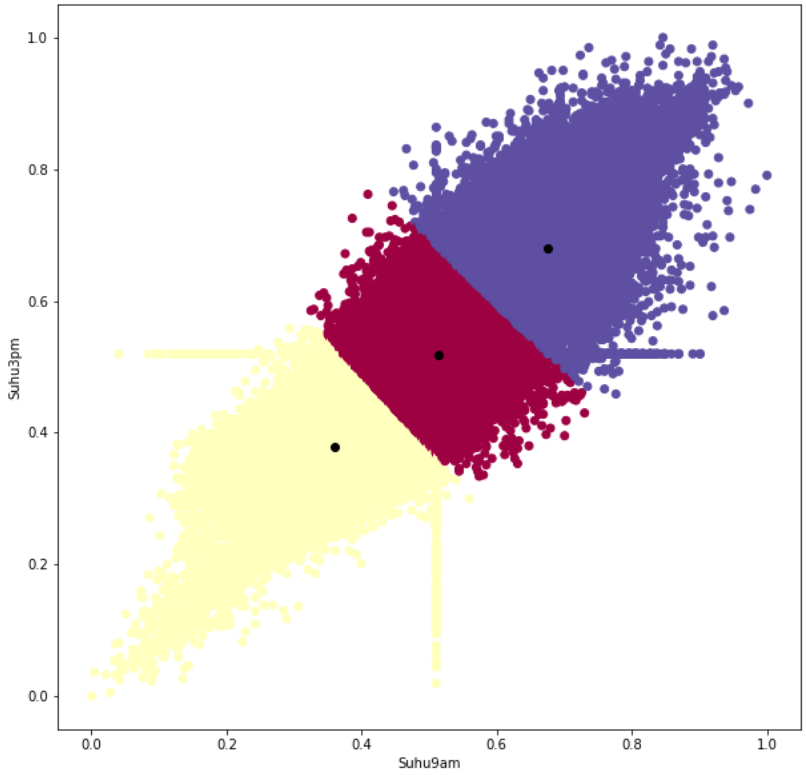
	SuhuMin	SuhuMax	Penguapan	Suhu9am	Suhu3pm
count	108510.000000	108510.000000	108510.000000	108510.000000	108510.000000
mean	0.488122	0.537716	0.035715	0.510370	0.519639
std	0.150304	0.136156	0.022009	0.136172	0.131566
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.379717	0.435701	0.027586	0.411392	0.424184
50%	0.483491	0.527831	0.033103	0.506329	0.512476
75%	0.596698	0.633397	0.035862	0.605485	0.606526
max	1.000000	1.000000	1.000000	1.000000	1.000000

3. Pemodelan

Terdapat beberapa fungsi yang dibuat dan juga terdapat sebuah main program untuk melakukan clustering, diantaranya yakni :

a. Fungsi *count_distance* yang digunakan untuk menghitung jarak antara suatu titik dengan titik yang lain secara garis lurus.
b. Fungsi *define_clusters* yang digunakan untuk mendefinisikan cluster dari masing-masing data yang ada dengan cara menghitung jarak dari data tersebut dengan centroid yang ada menggunakan fungsi *count_distance* yang sudah dibuat sebelumnya. Setelah mengetahui jarak dari sebuah data dengan masing-masing centroid yang ada, selanjutnya dimasukkanlah data tersebut kedalam cluster dengan jarak centroid terkecil.
c. Fungsi *count_centroids* yang digunakan untuk menentukan centroid baru dari sebuah cluster dengan cara menggunakan dataframe yang tiap data nya sudah masuk kedalam cluster tertentu lalu mencari mean atau rata-rata dari setiap cluster yang nantinya data tersebut akan menjadi centroid baru dari cluster tersebut.
d. Main Program menentukan berapa k (jumlah cluster) yang akan digunakan lalu sebelum melakukan loop mendefinisikan centroid terlebih dahulu secara acak dan juga menentukan cluster awal dari setiap data menggunakan centroid yang diacak tersebut. Selanjutnya setelah memiliki nilai awal untuk centroid dan juga cluster, menentukan berapa kali akan terjadi loop untuk mengulang-ulang fungsi *count_centroids* dan *define_clusters.* Diakhir, melakukan plot terhadap data dari clustering yang sudah dilakukan.

Hasil :

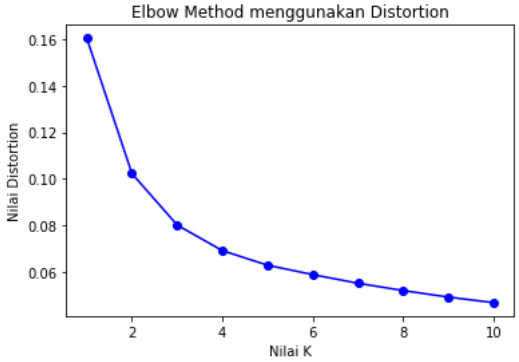


Setelah melakukan visualisasi data maka didapatkan plot hasil clustering seperti berikut dimana terdapat 3 cluster sesuai dengan nilai K yang sudah ditentukan sebelumnya.

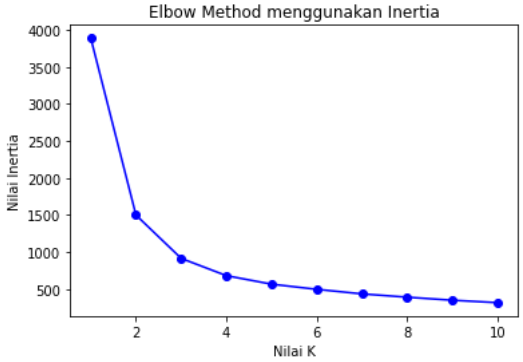
4. Evaluasi

Melakukan evaluasi dengan menggunakan *Elbow Method* yang memiliki 2 cara yakni menggunakan nilai distortion atau nilai inertia. Distortion adalah rata-rata jarak kuadrat dari titik tengah cluster dari masing-masing cluster. Inertia adalah jumlah dari nilai kuadrat jarak sampel ke titik tengah cluster terdekat. Dihitung untuk nilai K dari 1 sampai 10 agar dapat membuat grafik kelandaian antara nilai distortion/inertia terhadap nilai K. Nantinya akan ditentukan nilai K terbaik berdasarkan grafik tersebut yang dimana ditentukan dengan cara memilih nilai K yang merupakan titik dimana nilai distortion/inertia dari data tersebut mulai menurun secara linear.

Grafik menggunakan Distortion :

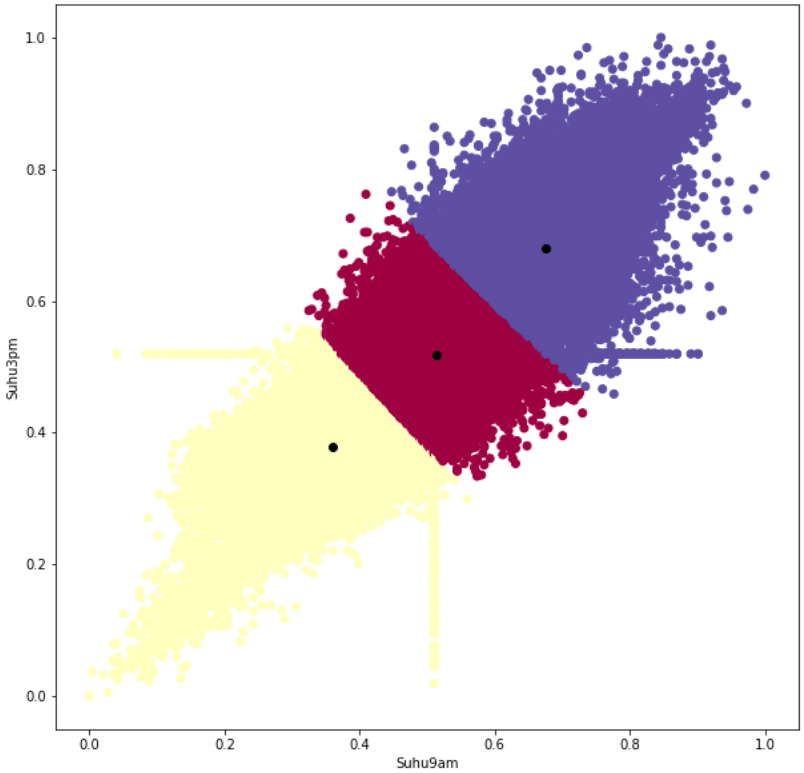


Grafik menggunakan Inertia :

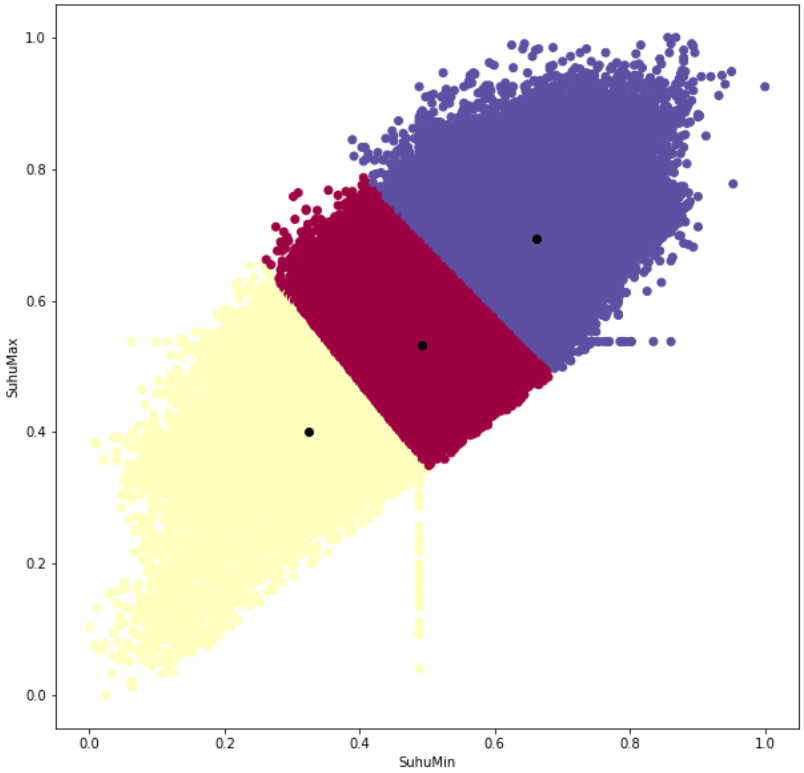


Dari 2 grafik tersebut, didapatkan hasil bahwa nilai K yang terbaik adalah 3 dikarenakan setelah nilai K = 3, grafik mulai menurun secara linear.

5. Eksperimen
    a. Melakukan clustering untuk kolom yang berbeda
       Kolom Suhu9am dan Suhu3pm :



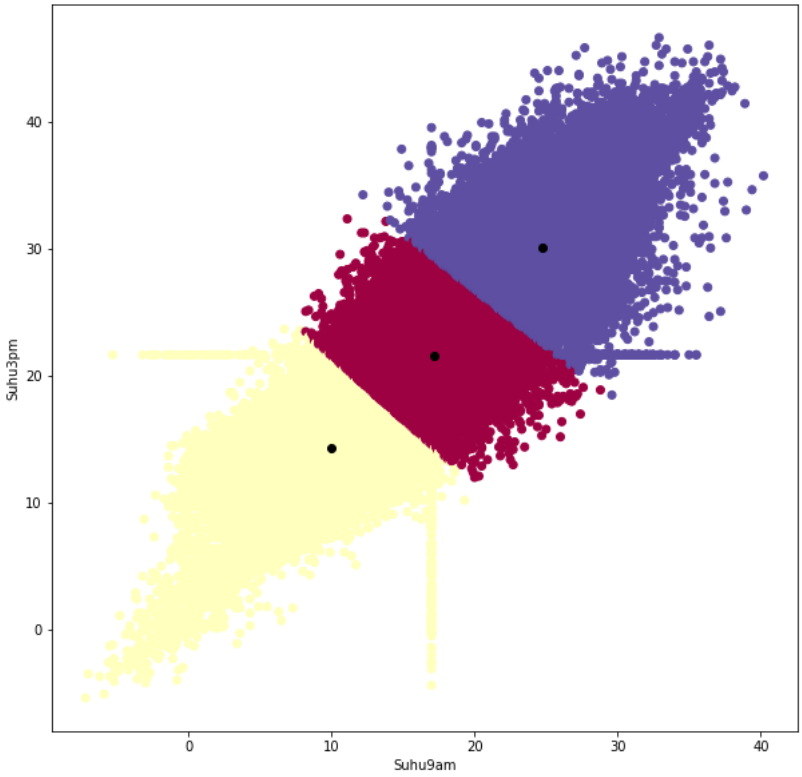
Kolom SuhuMin dan SuhuMax :



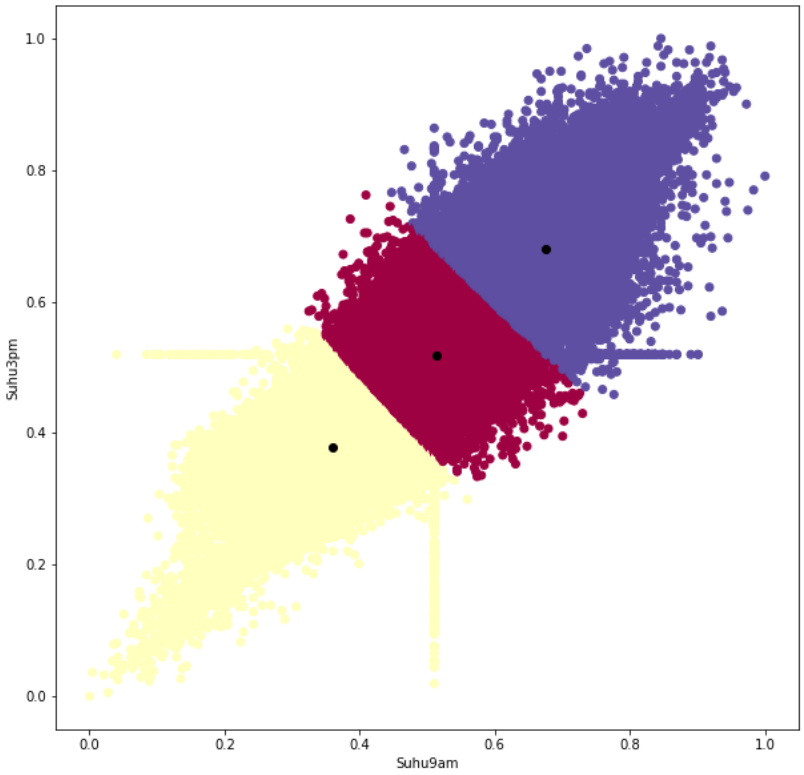
Jika menggunakan kolom yang berbeda, maka plot clustering akan berbeda dikarenakan nilai pada kolom yang berbeda.

b. Melakukan clustering untuk data yang sudah dinormalisasi dan belum :

Sebelum normalisasi :



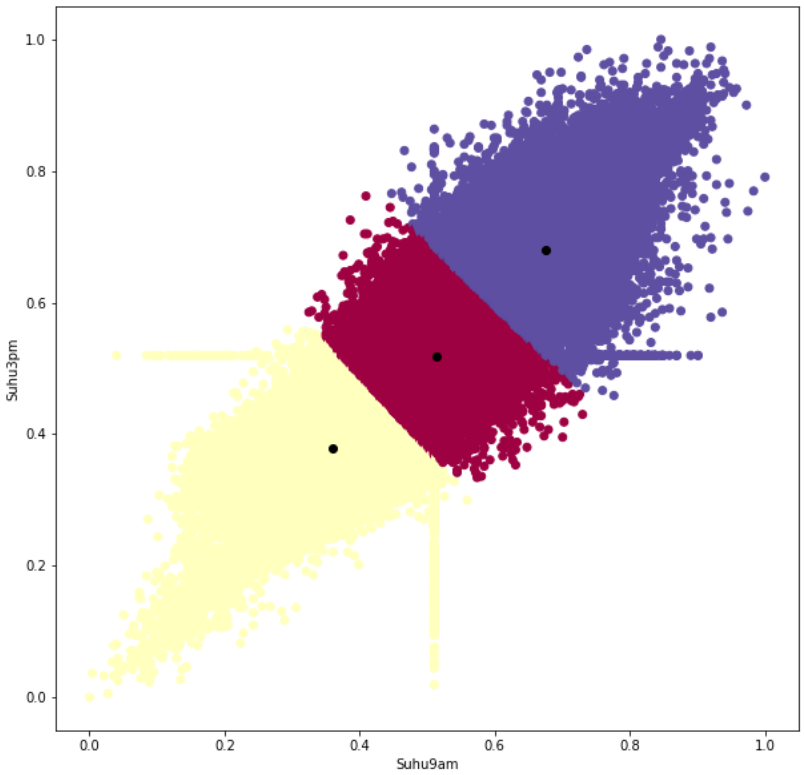
Setelah normalisasi :



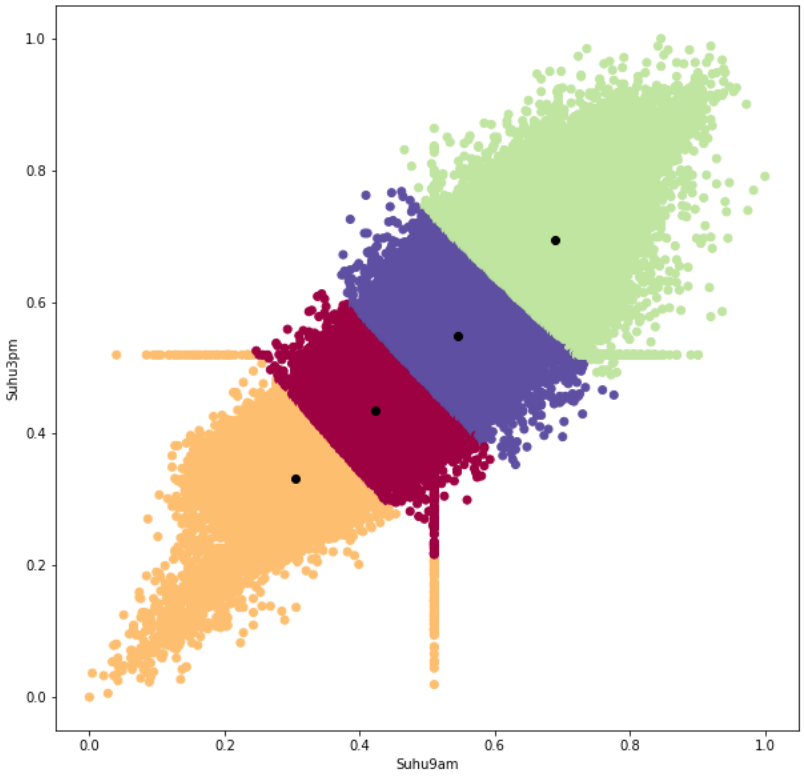
Jika melakukan clustering dengan atau tanpa normalisasi, maka yang berbeda hanayalah untuk range nilai sebelum normalisasi masih besar.

c. Melakukan clustering untuk nilai K yang berbeda

K = 3 :



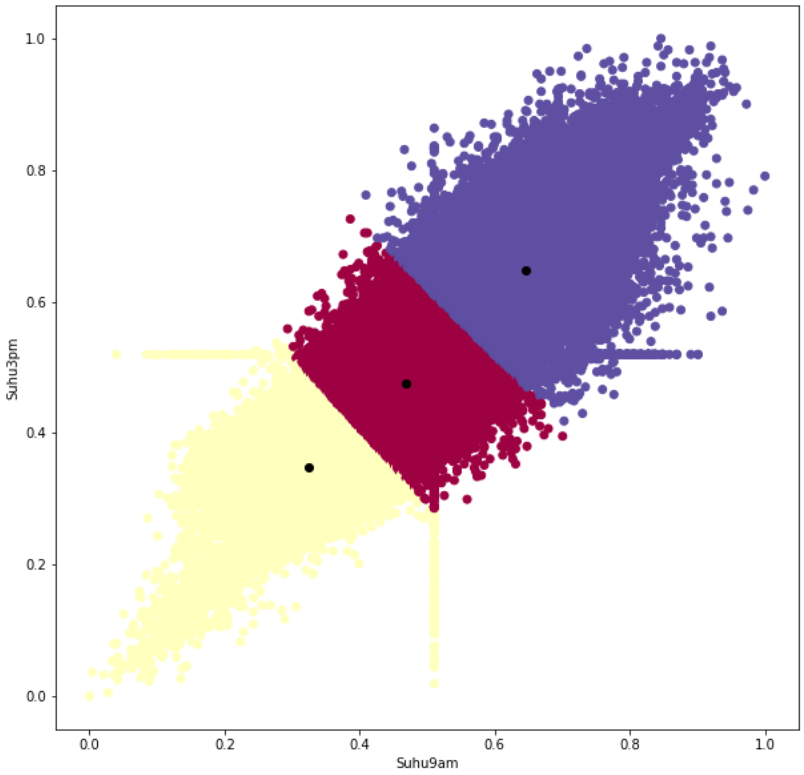
K = 4 :



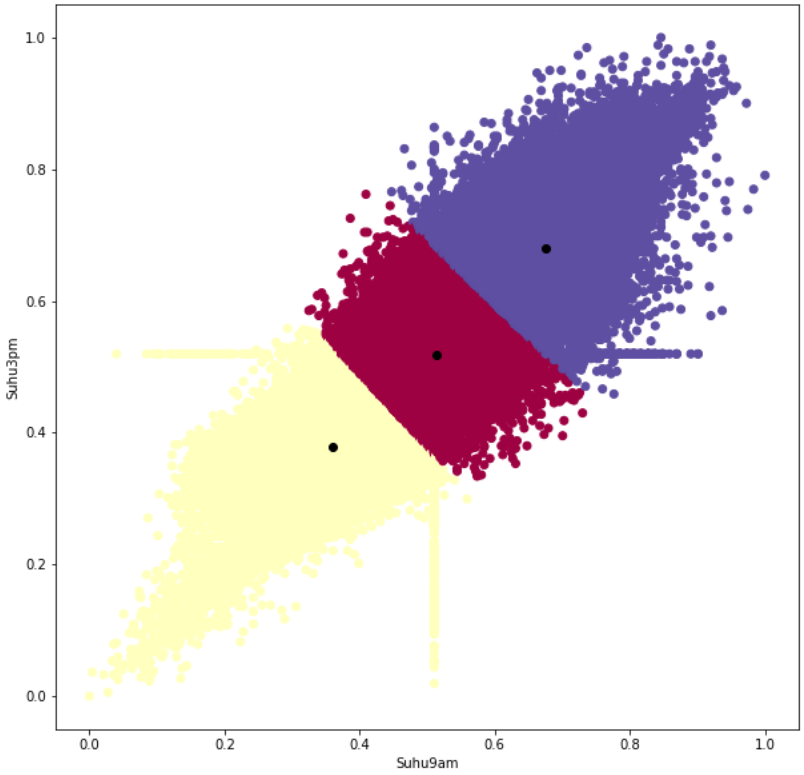
Jika melakukan clustering dengan K yang berbeda, maka akan terdapat perbedaan pada jumlah cluster yang ada dan juga ada sedikit perbedaan untuk cluster dari masing-masing data.

d. Melakukan clustering untuk jumlah iterasi yang berbeda

Iterasi = 10 :



Iterasi = 20 :



Jika menggunakan jumlah iterasi yang berbeda, maka clustering akan cenderung lebih random ketika iterasinya sedikit dan akan cenderung sama berapa kali pun dicoba ketika iterasinya semakin banyak.

6. Kesimpulan

Untuk melakukan clustering perlu untuk melakukan beberapa tahap diantaranya yakni :

a. Pra-Pemrosesan Data

Dalam proses ini, data diolah sehingga tidak memiliki nilai duplikat, tidak memiliki nilai null dan dinormalisasi agar range nilai tiap kolom berkisar antara 0 sampai 1.

b. Pemodelan

Dalam proses ini, dilakukan random untuk nilai centroid sebanyak k yang sudah ditentukan dan selanjutnya dilakukan penentuan cluster untuk setiap data yang akan diiterasi sebanyak jumlah iterasi yang sudah ditentukan.

c. Evaluasi

Dalam proses ini, dilakukan perhitungan nilai distortion dan inertia untuk mengoptimasi dan menentukan nilai K terbaik.