Nama : **Shidqi Aqil Naufal**

Nim : **1301164147**

Kelas : **IF 40-08**

TUGAS 2

1. **Kelebihan k-Means Clustering :**
   a. Menggunakan prinsip yang sederhana, dapat dijelaskan dalam non-statistik

   b. Waktu yang dibutuhkan untuk menjalankan nya relatif cepat

   c. Sangat fleksibel, dapat dengan mudah diadaptasi.

   Kekurangan k-Means Clustering :

   a. Karena menggunakan k buah acak, tidak di jamin untuk menemukan kumpulan cluster yang optimal
   b. dapat terjadinya curse of dimensionality, apabila jarak antara cluster yang satu dengan yang lain memiliki banyak dimesi.

   Contoh Kasus : Algoritma k-means dapat diterapkan pada permasalahan koleksi perpustakaan. Koleksi perpustakaan merupakan salah satu indicator untuk enentukan kualitas dari sebuah perpustakaan. Semakin lengkap koleksinya dan semakin memenuhii kebuthuan pengunjungnya maka semakin menarik. Untuk itu dapat digunakan metode klasterisasi k-means. Data yang digunakan dapat berupa data umur, gender, pekerjaan, dan Pendidikan pengunjung untuk menentukan koleksi apasaja yang diminati oleh pengunjung sehingga membantu dalam penentuan kebijakan pengadaan koleksi.

2. Agglomerative Hierarchical Clustering merupakan salah satu metode pada hierarchical clustering yang mana konsep dasar dari metode ini adalah dengan meletakkan setiap objek sebagai sebuah cluster tersendiri. Setelah menjadikan setiap objek menjadi cluster maka akan digabungkan menjadi cluster yang lebih besar sampai akhirnya menyatu dalam sebuah cluster. Ada beberapa metode untuk menukur kedekatan dalam agglomerative yaitu single link, multi link, group average, mean.

   Contoh Kasus : Penggunaan metode *Agglomerative Hierarchical Clustering* dapat digunakan dalam mengatasi besarnya dimensi dalam data microarray. Data microarray adalah data yang dihasilkan dari teknologi microarray yang mana teknologi tersebut akan menghasilkan data ekspresi gen manusia. Dalam data microarraty mempunyai permasalahan besarnya dimensi yang dihasilkan. Masalah ini dapat diatasi dengan menggunakan metode *Agglomerative Hierarchical Clustering* yang mana dengan menggunakan metode ini akan dihasilkan sekumpulan objek yang mempunyai kesamaan tertentu dalam satu klaster.

Laporan Klasterisasi Data dengan Menggunakan Metode Self Organizing Map (SOM)

Deskripsi Masalah

Disediakan suatu dataset yang memiliki 600 sampel data dan dua atribut. Data yang diberikan tidak diberikan label kelas. Diperlukan sebuah model sistem klasterisasi untuk mengelompokkan data tersebut kedalam kelompok tertentu. Metode yang digunakan untuk klasterisasi dataset ini adalah *Self Organizing Map* (SOM) dengan mencari jumlah klaster yang palin optimum.

Metode Penyelesaian

Metode yang akan digunakan adalah metode ***Self Organizing Map*** **(SOM).** Dalam metode ini berfokus pada pencarian nilai atau letak neuron yang nantinya akan dihitung kedekatannya setiap neuron pada setiap data sampel yang ada. Adapun cara kerja metode ***Self Organizing Map*** **(SOM)** sebagai berikut **:**

1. Tentukan jumlah neuron yang akan digunakan. Jumlah neuron yang digunakan akan menentukan jumlah klaster yang akan dibuat.

2. Generate secara random atribut (letak) dari neuron. Karena data sampel yang diberikan mempunyai dua atribut, maka atribut yang akan di generate akan mempunyai dua atribut pula. Generate random akan dilakukan dalam range 0-17, karena atribut pada data sampel mempunyai rentang batas bawah tidak kurang 0 dan batas atas tidak lebih dari 17.

3. Hitung jarak antara sample data dengan setiap neuron yang telah di generate. Lakukan perhitungan jarak dengan Euclidean distance yang mempunyai persamaan sebagai berikut:

   $$S = \sqrt{\Sigma(xi - yi)^2}$$

   dimana xi adalah atribut pada data dan yi pada neuron.

4. setelah itu, cari jarak antara objek dan neuron yang paling dekat.

5. Kemudian, lakukan perhitungan T, yang mana mempunyai persamaan sebagai berikut:

   $$T_{j,I(\mathrm{x})} = \exp\left(-S_{j,I(\mathrm{x})}^2 / 2\sigma^2\right)$$

   Dimana S adalah jarak yang dicari pada step 3, yaitu Euclidean distance, dan σ adalah sigma, yang mana nilai sigma akan selalu berubah pada setiap iterasi yang dilakukan. Inisiasi sigma diawal adalah 2.

6. setelah itu, lakukan perhitungan Δw. Dalam perhitungan ini, akan posisi neuron baru. Adapun persamaannya seperti berikut:

   $$\Delta w_{ji} = \eta(t) \cdot T_{j,I(\mathrm{x})}(t) \cdot \left(x_i - w_{ji}\right)$$

Dimana n(t) merupakan learning rate pada iterasi tertentu. Pada kasus ini, inisiasi awal learning rate adalah 0.1.

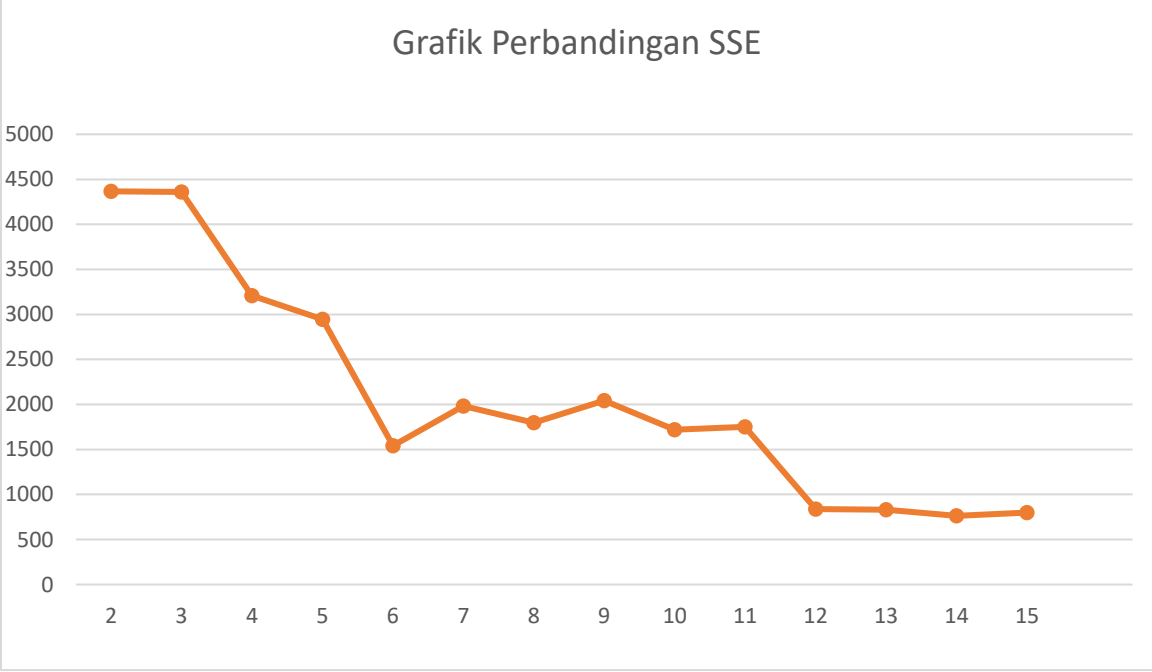
7. setelah melakukan hingga step 6, maka satu iterasi telah dilakukan. Dalam kasus ini, iterasi yang dilakukan adalah sebanak 100 kali. Dimana setiap iterasi, akan meng update sigma dan learning ratenya dengan persamaan seperti berikut :

$$\sigma(t) = \sigma_0 \exp(-t/\tau_\sigma)$$

$$\eta(t) = \eta_0 \exp(-t/\tau_\eta)$$

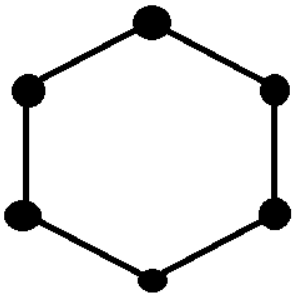
8. setelah semua iterasi dilakukan, maka lakukan perhitungan jarak kembali untuk setiap sampel data dengan hasil neuron iterasi terakhir. Misalnya jarak terdekat sampel data pertama adalah dengan neuron 6, maka sampel data pertama akan masuk kedalam klaster 6.
9. lakukan pengecekan SSE, yang mana bertujuan untuk mencari jumlah neuron paling optimal. Hasil rentang hasil SSE yang signifikan antara setiap neuron adalah yang nantinya akan dipilih menjadi jumlah neuron yang paling optimum.

Hasil Pemrosesan



Hasil pemorsesan klasterisasi dengan metode SOM adlah setiap data yang diberikan telah dikelompokkan menjadi 6 klaster. Pemilihan 6 klaster ini berdasarkan SSE yang telah dicari di rentang neuron 2 sampai 15. Klaster 6 dipilih karena perbedaan SSE yang paling signifikan dari neuron sebelumnya.

Hasil klasterisasi ini pula ditentukan dengan susunan neuron yang dibangun. Dalam implementasi ini neuron yang dibangun adalah dengan menggunakan dua dimensi. Jadi setiap neuron hanya akan mempunyai dua tetangga terdekatnya.



Gambar diatas adalah gambaran dari 6 neuron yang dibangun. Hal ini sangat memengaruhi bagaimana metode SOM bekerja. Apabila neuron disusun 3 dimensi, maka hasilnya akan berbeda pula.

Hasil dari proses klasterisasi data dapat dilihat pada gambar dibawah ini :

1	9.802	10.132	4
2	10.35	9.768	0
3	10.098	9.988	2
4	9.73	9.91	2
5	9.754	10.43	4
6	9.836	9.902	2
7	10.238	9.866	0
8	9.53	9.862	2
9	10.154	9.82	0
10	9.336	10.456	4
11	9.378	10.21	2
12	9.712	10.264	4
13	9.638	10.208	4
14	9.518	9.956	2
15	10.236	9.91	0
16	9.4	10.086	2
17	10.196	9.746	0
18	10.138	9.828	0
19	10.062	10.26	4
20	10.394	9.984	0

Gambar diatas adalah gambaran output yang dihasilkan dari model sistem yang dibuat. Yang mana pada kolom satu dan dua adalah atribut dari data dan pada kolom ketiga adalah letak klaster. Jadi misalkan data pada sampel pertama, mempunyai atribut 9.802 dan 10.132, maka masuk kedalam klaster 4.

Hasil lebih lengkap telah dilampirkan pada file Excel.