KECERDASAN KOMPUTASIONAL

TUGAS 1

Kelas: B

Anisa Aurafitri	05111840000049
Afia Hana Yusriya	05111840000111
Ivan Abdillah Rahman	05111840000137

Dosen: Dr. Diana Purwitasari, S.Kom., M.Sc.

DEPARTEMEN TEKNIK INFORMATIKA

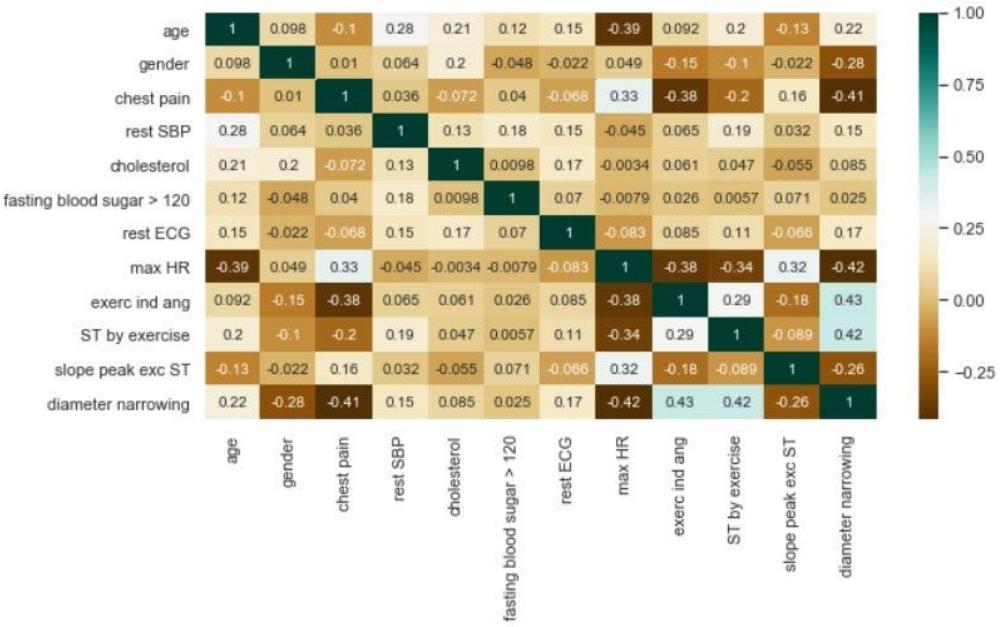
FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

2020

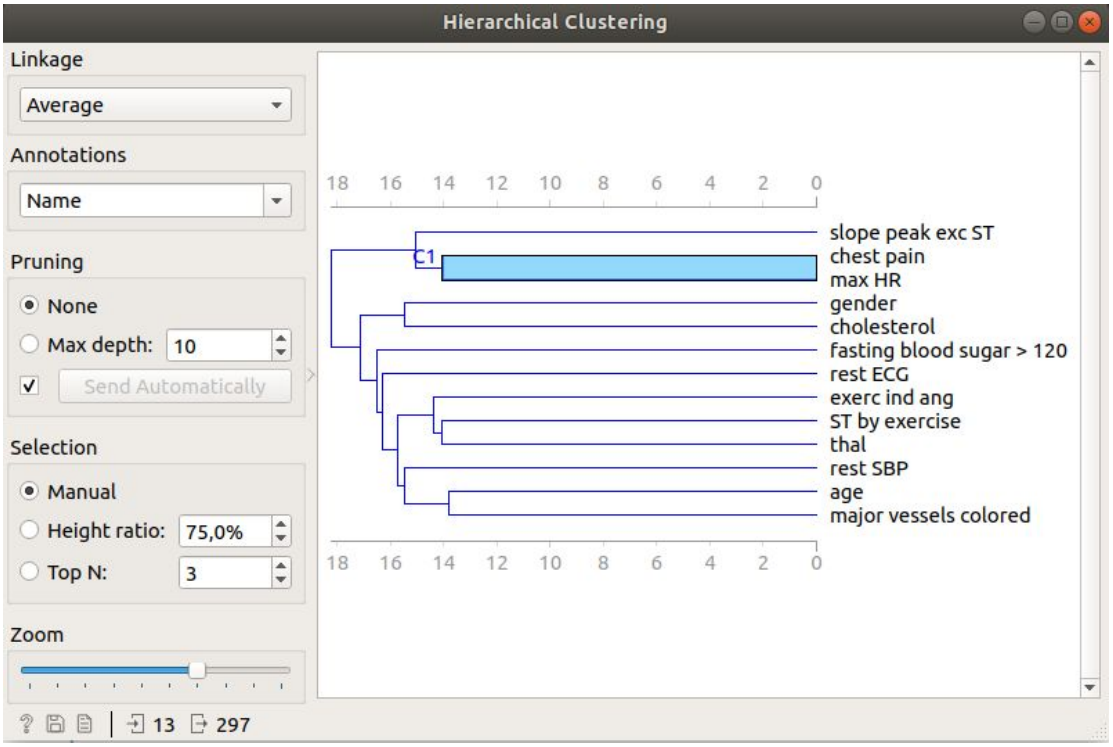
1. **Apa kombinasi fitur yang memberikan cluster terbaik menurut indikator Silhouette score?**

   Kami menggunakan data set “heart_disease.tab”.
   Kami mengubah data categorical menjadi numeric terlebih dahulu, kemudian dicari heatmapnya dan didapat seperti gambar dibawah.

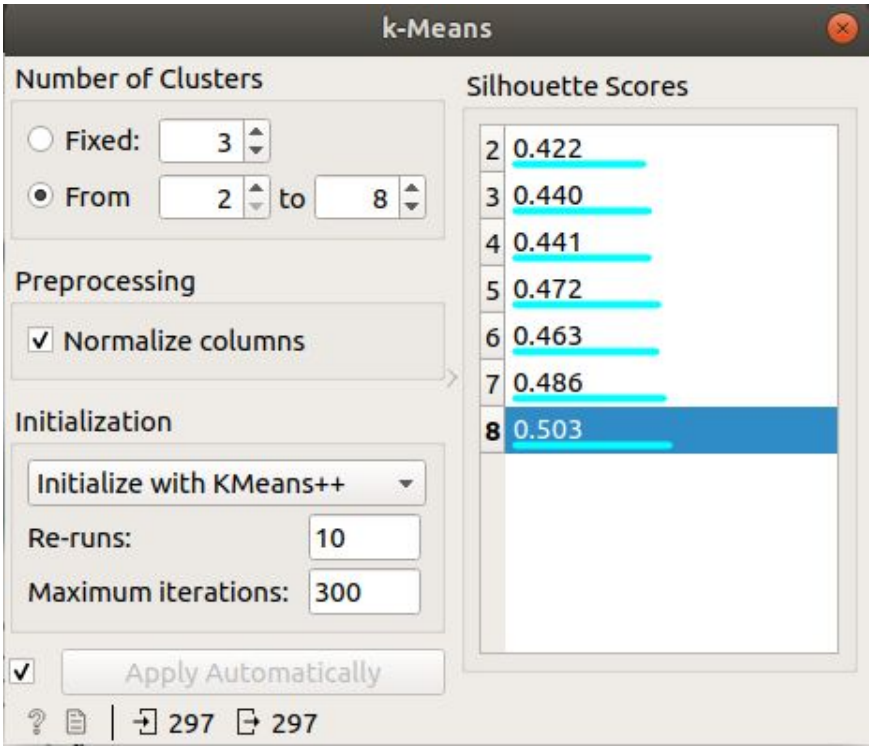


   Dari heatmap tersebut didapat nilai tertinggi 0.33 pada kombinasi *chest pain* dan *max HR*. Meskipun terdapat angka 0.42 dan 0.43, angka tersebut tidak digunakan karena *diameter narrowing* merupakan target dari data ini.

   Kemudian dapat dilihat juga menggunakan hierarchical clustering dibawah ini, dapat dilihat bahwa *chest-pain* dan *max HR* menjadi satu cluster terdekat.



Silhouette score antara *chest-pain* dan *max HR* juga menunjukkan nilai yang cukup tinggi daripada kombinasi lainnya.



Sehingga kami menyimpulkan bahwa kombinasi yang memberikan cluster terbaik yaitu *chest pain* dan *max HR*.

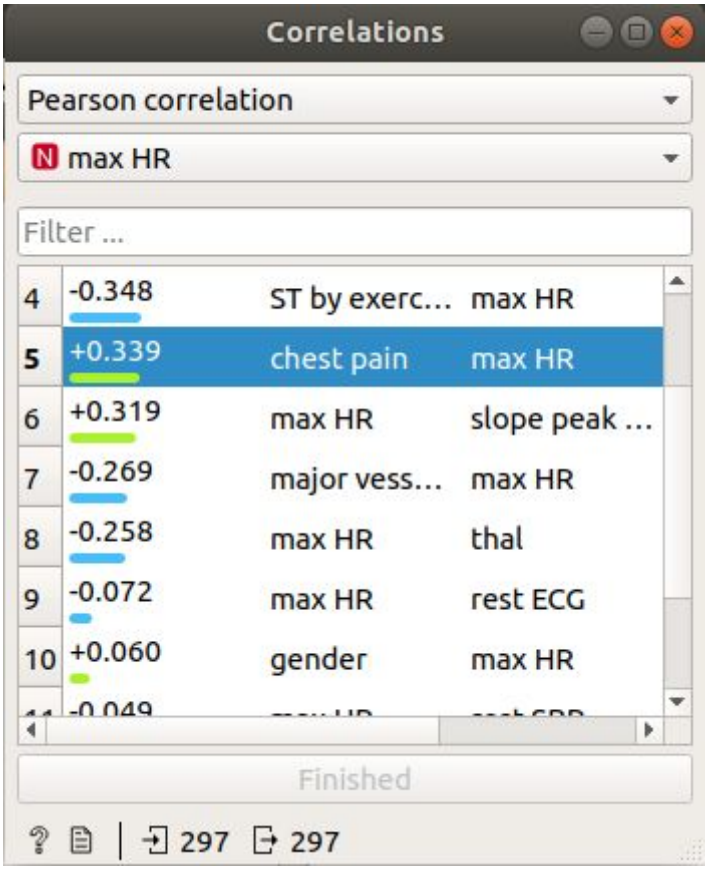
2. **Bagaimana hubungan korelasi antar fitur-fitur yang terpilih?**

Menurut data “heart_disease.tab”, chest-pain yang dimaksud adalah Angina.

Angina adalah sakit di dada (chest-pain) yang disebabkan oleh jantung yang kurang mendapat asupan oksigen yang cukup.

Max HR (*Maximum Heart Rate*) adalah jumlah maksimum detak jantung yang dilakukan oleh jantung kita selama 1 menit ketika melakukan sesuatu. *Maximum heart rate* dapat dihitung dengan cara mengurangi sebanyak umur kita dari 220. Misal sekarang kita 20 tahun, maka *maximum heart rate* kita adalah 220-20 = 200 MHR. Berarti 200 adalah jumlah maksimum jantung kita bekerja saat olahraga atau melakukan pekerjaan berat lainnya.

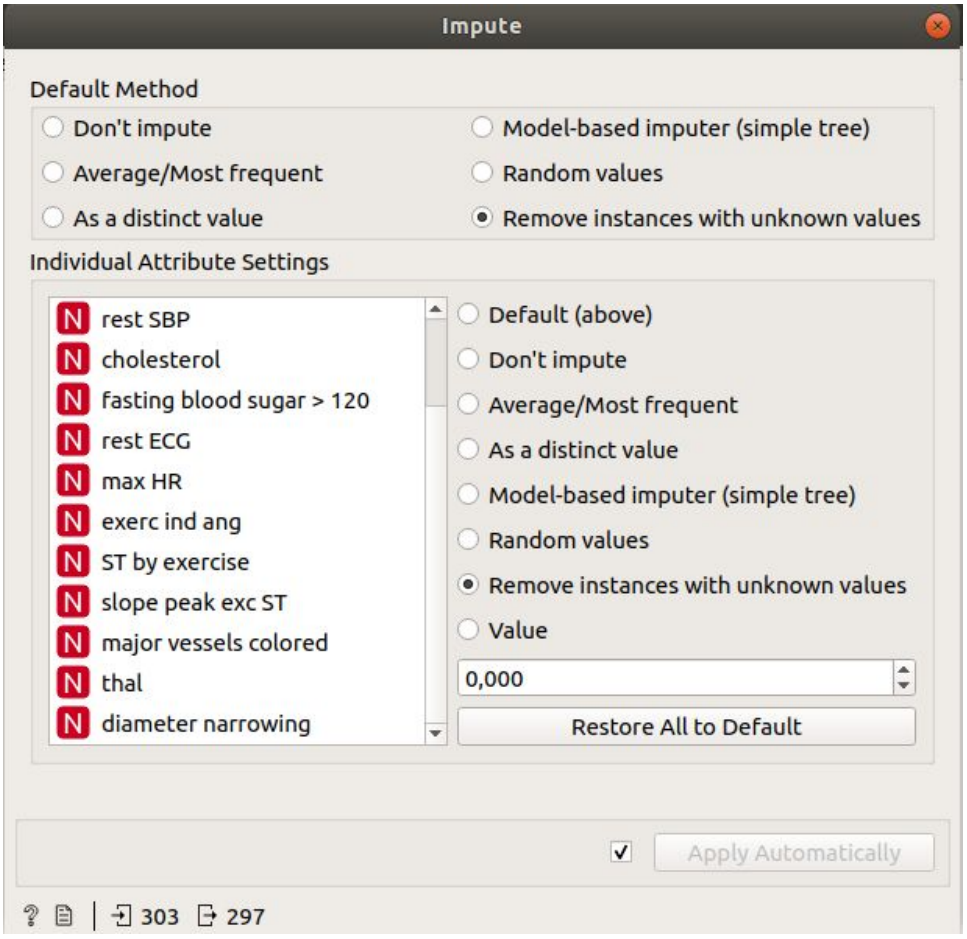
Ketika *maximum heart rate* seseorang melebihi nilai normal, maka dapat memicu terjadinya *chest-pain* (sakit/rasa tidak nyaman di daerah dada).



Selain itu, angka korelasi antara *chest-pain* dan *max HR* juga cukup tinggi.

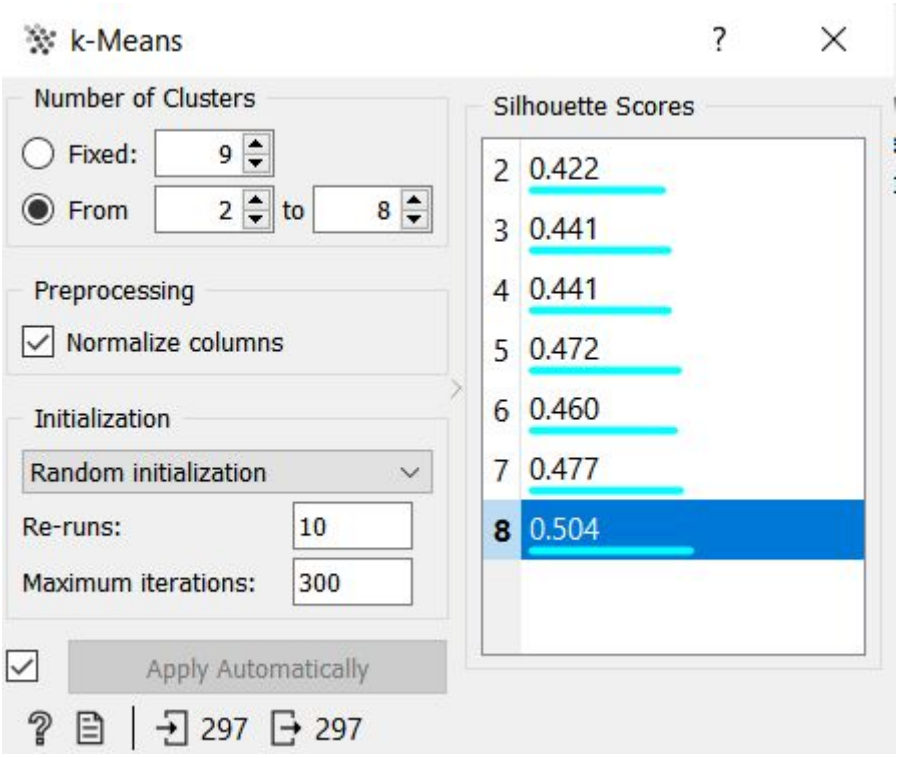
3. **Apakah ada proses cleaning data (impute, dll)?**

   Ada. Ketika melakukan impute pada data, dapat dilihat pada bagian bawah yang awalnya terdapat 303 data menjadi 297 data. Sehingga terdapat 6 data yang memiliki null value.

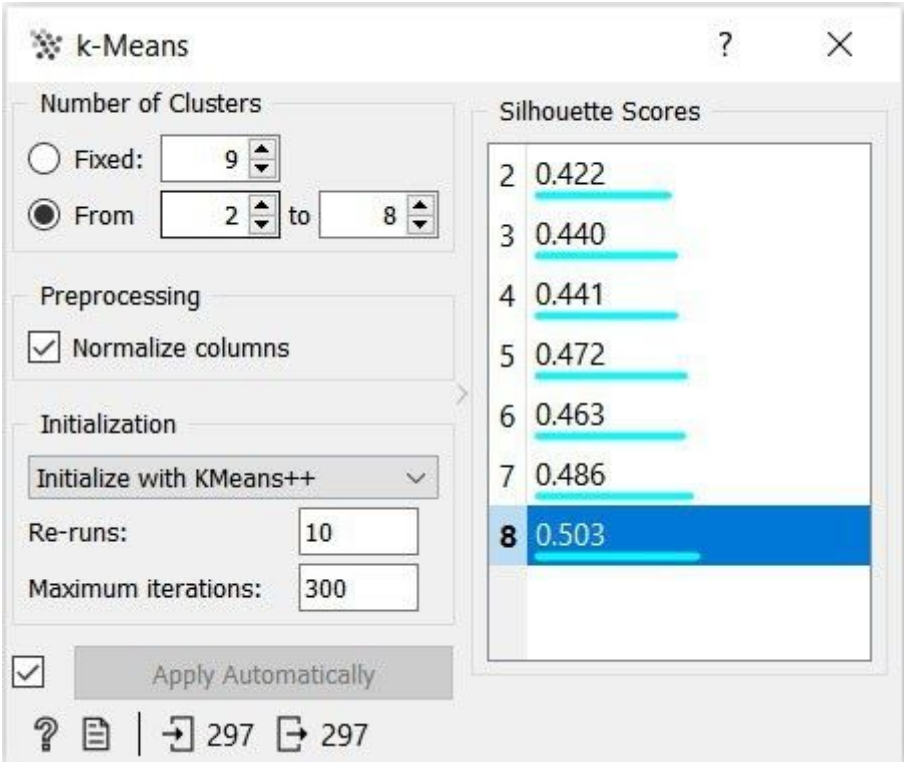


4. **Bagaimana hasil cluster jika centroid terbentuk secara random atau dipiihkan yang terbaik K-Means ++**

   Pada kombinasi fitur *chest pain* - *max HR*, jumlah cluster tidak berubah tetapi merubah silhouette scores. Dimana pada inisialisasi random memiliki score 0.504 sedangkan pada k-means++ memiliki score 0.503.



Inisialisasi diatas menggunakan inisialisasi random menunjukkan hasil terbaik pada cluster = 8 dan silhouette score = 0.504

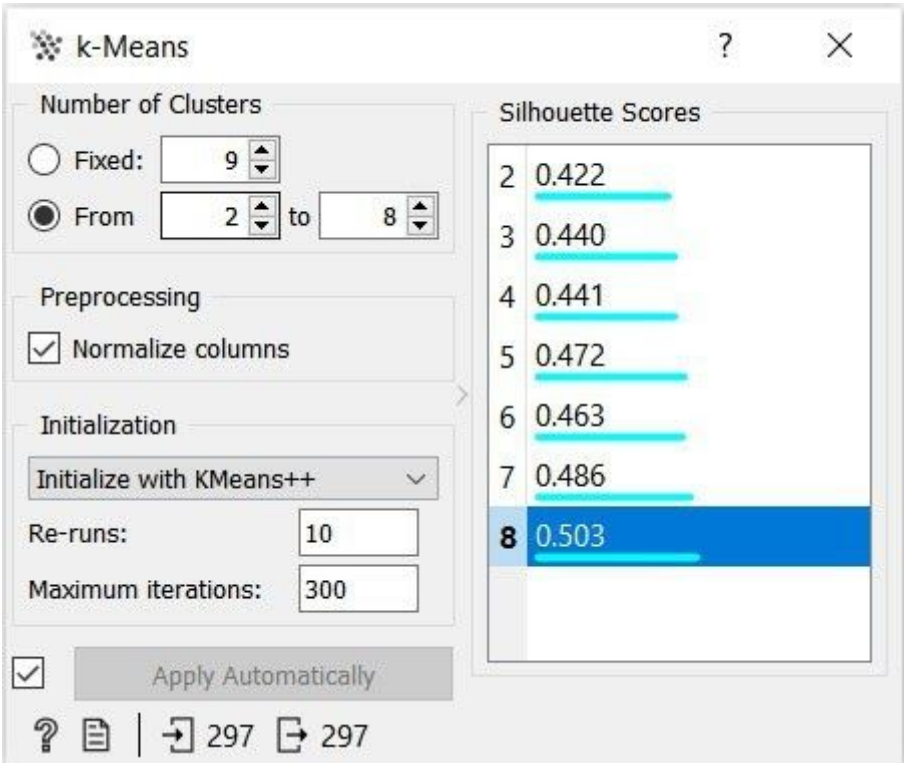


Inisialisasi diatas menggunakan inisilisasi k-means++ menunjukkan hasil terbaik pada cluster = 8 dan silhouette score = 0.503

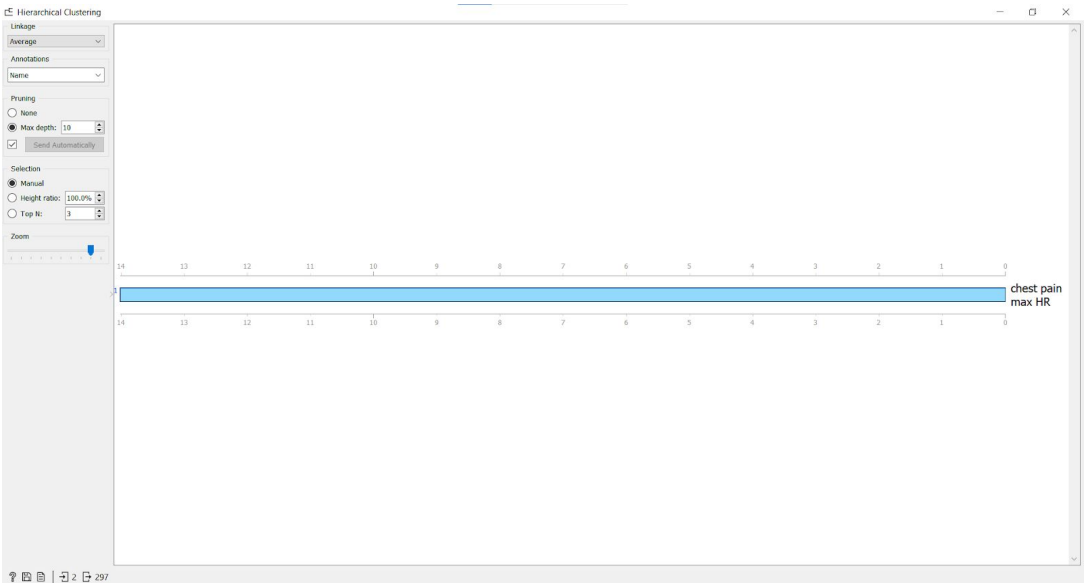
5. **Bandingkan K-Means dengan Hierarchical (divisive, agglomerative)**

K-Means: Saat melakukan K-Means, kami bisa langsung memilih banyak fitur untuk di clustering. Disini kami menggunakan 2 fitur dan mendapatkan bahwa kami mendapat hasil paling baik pada cluster=2 dan silhoutte score = 0.503.

Hierarchical: Saat melakukan Hierarchical, kami harus memilih fitur untuk ditentukan cluster secara satu per satu. Dan untuk menentukan jumlah akhir cluster harus ditentukan secara manual.



Menggunakan K-Means

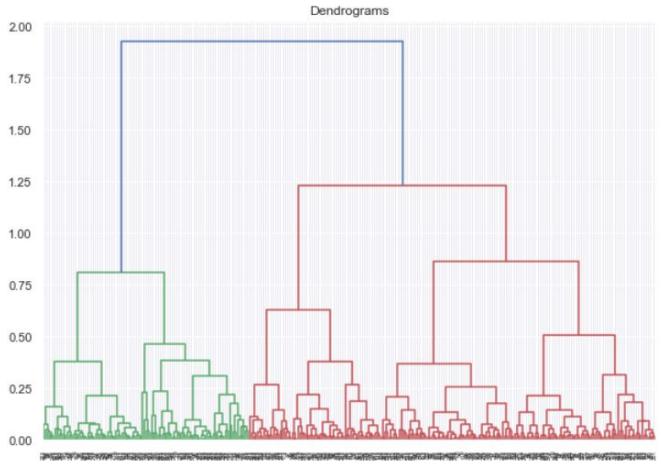


Menggunakan Hierarchical

6. **Bandingkan dengan algoritma jika menggunakan SkLearn**
   Saat menggunakan SKLearn untuk Hierarchical

```
In [57]: plt.figure(figsize=(10, 7))
         plt.title("Dendrograms")
         dend = shc.dendrogram(shc.linkage(data_scaled, method='ward'))
         plt.axhline(y=6, color='r', linestyle='--')
```

```
Out[57]: <matplotlib.lines.Line2D at 0x1ac92f3ea88>
```



Untuk kodingan lengkap bisa dicek di github.

7. **Kesimpulan**
   Untuk clustering data **heart disease**, jika ditinjau dari segi kecepatan proses maka baik KMeans maupun Hierarchical hampir sama cepatnya karena jumlah data heart disease sendiri tergolong tidak banyak. Tetapi menurut kami menggunakan KMeans lebih baik karena mendapat jumlah cluster yang jelas dengan fitur yang jelas juga, serta dapat memilih lebih dari 1 fitur. Sedangkan jika menggunakan Hierarchical untuk fitur harus diuji satu-satu

lalu kita memilih 1 fitur saja.