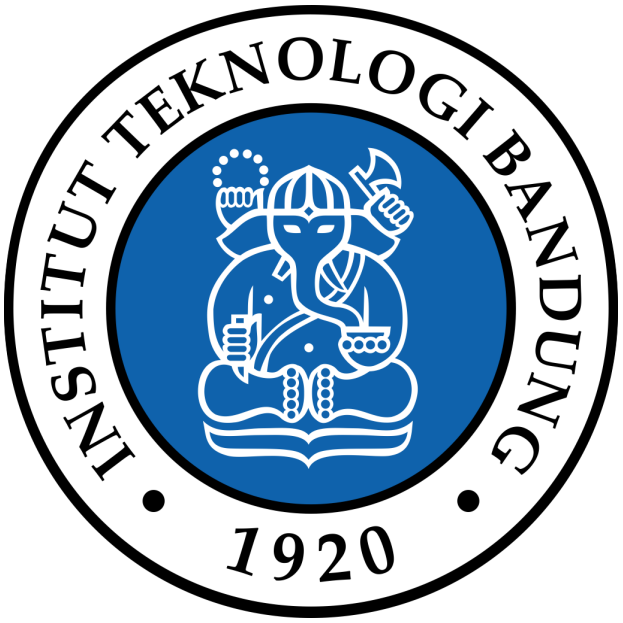
Laporan Tugas Besar II

IF3270 Pembelajaran Mesin



13517029 Reyhan Naufal Hakim

13517035 Hilmi Naufal Yafie

13517095 Naufal Zhafran Latif

13517098 Anzaldi Sulaiman Oemar

PROGRAM STUDI TEKNIK INFORMATIKA

SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

2020

A. Implementasi K-Means

KMeans merupakan salah satu algoritma clustering yang cukup mudah untuk diimplementasikan. Pada modul MyKMeans, implementasi menggunakan modul numpy untuk mempermudah dan membuat kode terlihat lebih bersih.

Diawal akan diterima input berupa jumlah cluster yang akan dibentuk yang pada algoritma ini biasa disimbolkan dengan variabel k. Selain itu input yang diterima adalah berupa sebuah array yang berisi data-data yang akan digunakan untuk clustering. Setelah semua input sudah terpenuhi maka MyKMeans dapat melakukan clustering

Pada tahap clustering, dimulai dengan memilih secara random centroid-centroid awal untuk setiap clusternya. Jumlah centroid yang dipilih berdasarkan nilai k yang telah didefinisikan diawal. Setelah centroid, setiap data akan ditentukan clusternya masing-masing. Penentuan cluster ini dilakukan dengan mencari nilai terendah dari jarak sebuah data ke cluster. Jarak menggunakan rumus jarak euclidean. Setelah setiap data sudah dikelompokkan berdasarkan clusternya masing-masing, setiap cluster harus melakukan update pada centroidnya dengan menghitung rata-rata dari setiap data pada cluster. Dari rata-rata tersebut akan dihasilkan nilai centroid yang baru. Proses ini kembali diulang mulai dari penetuan cluster untuk setiap data. Proses iterasi dilakukan hingga centroid tidak ada perubahan atau perubahan sudah berada pada suatu threshold yang ditentukan

B. Implementasi Agglomerative

Pada MyAgglomerative, pertama-tama dibuat clustering tree berdasarkan jenis linkage yang dipilih (single, complete, average, atau average group/centroid). Pembuatan clustering tree dengan cara mengasumsikan setiap data pada dataset yang diberikan merupakan sebuah cluster tersendiri, lalu dari tiap cluster dicari pasangan cluster dengan jarak minimum sesuai linkage yang dipilih untuk dijadikan sebuah cluster baru, hingga akhirnya membentuk sebuah cluster yang berisi seluruh dataset. Setiap cluster yang terbentuk akan disimpan dengan nilai cluster terakhir ditambah satu. Sebagai contoh, semisal terdapat sepuluh cluster diawal, dan cluster satu dan cluster dua yang digabungkan menjadi cluster, maka cluster ini akan disimpan sebagai cluster sebelas, dan pada iterasi berikutnya, penggabungan cluster akan disimpan sebagai cluster dua belas, begitu seterusnya hingga terbentuk satu cluster utuh. Untuk membentuk n buah cluster, maka dilakukan pencarian dari cluster paling akhir terbentuk hingga mencapai n buah cluster. Lalu untuk setiap cluster, dilakukan pencarian hingga daun dari cluster tersebut untuk diberi label yang sama. Nilai daun disini adalah urutan data pada dataset, sehingga nantinya akan menghasilkan array yang berisi label clustering, dimana pada arrray ke-i, merupakan representasi dari data ke-i pada dataset.

Pemilihan linkage yang digunakan pada MyAgglomerative akan sangat berpengaruh pada hasil clustering, karena nilai jarak yang dihasilkan dari setiap linkage dapat berbeda. Pada single linkage, jarak antar cluster ditentukan dengan jarak terpendek dari setiap kemungkinan jarak data antar cluster. Lalu, pada complete linkage, jarak antar cluster ditentukan dengan jarak terjauh dari setiap kemungkinan jarak data antar cluster. Sementara itu, pada average linkage, jarak antar cluster ditentukan dengan menghitung rata-rata dari setiap kemungkinan jarak data antar cluster. Dan pada average group linkage, jarak antar cluster ditentukan dengan mencari rata-rata pada masing-masing cluster (centroid) terlebih dahulu, lalu menghitung jaraknya, sehingga linkage ini disebut juga centroid linkage.

Sebelum melakukan clustering pada dataset iris, dataset terlebih dulu dipisahkan dengan labelnya. Untuk hal ini, digunakan fungsi bawaan dari sklearn, yaitu load_iris() ke sebuah variabel, lalu gunakan variabel.data untuk menggunakan dataset iris.

C. Evaluasi

Evaluasi dilakukan dengan perhitungan metrik fowlkes-mallows dan silhouette coefficient yang memanfaatkan fungsi bawaan dari library sklearn pada bahasa python. Metode evaluasi pertama adalah fowlkes-mallows score. Menurut definisi, nilai fowlkes-mallows merupakan evaluasi kemiripan klaster yang dihitung dengan memanfaatkan rata-rata geometrik dari kesamaan pasangan prediksi klaster. Fowlkes-mallows memiliki nilai antara 0 sampai 1, semakin tinggi nilai fowlkes-mallows menandakan metode yang digunakan untuk klasterisasi dari pasangan-pasangan tersebut semakin mirip.

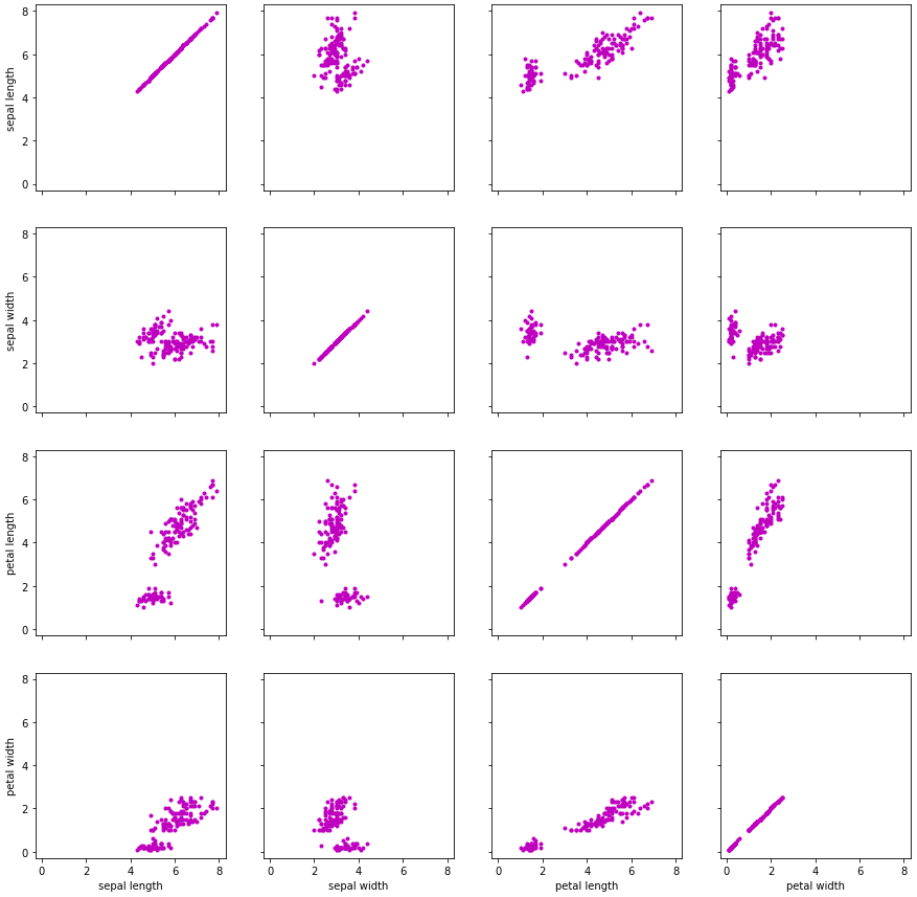
Metode evaluasi kedua adalah silhouette coefficient, metode ini mengukur kedekatan antara semua klaster, klasterisasi dikatakan baik apabila masing-masing klaster yang terbentuk memiliki jarak perbedaan yang sangat jauh dan jelas terpisah. Silhouette coefficient akan menghasilkan nilai antara -1 sampai 1. Rangkuman hasil evaluasi dari masing-masing algoritma tertulis dalam tabel berikut.

Algoritma	Fowlkes-Mallows Score	Silhouette Coefficient Score
K-Means	0.8208	0.5528
Agglomerative (Single linkage)	0.7635	0.5121
Agglomerative (Complete linkage)	0.7635	0.5121
Agglomerative (Average linkage)	0.8407	0.5542
Agglomerative (Average group linkage)	0.8407	0.5542

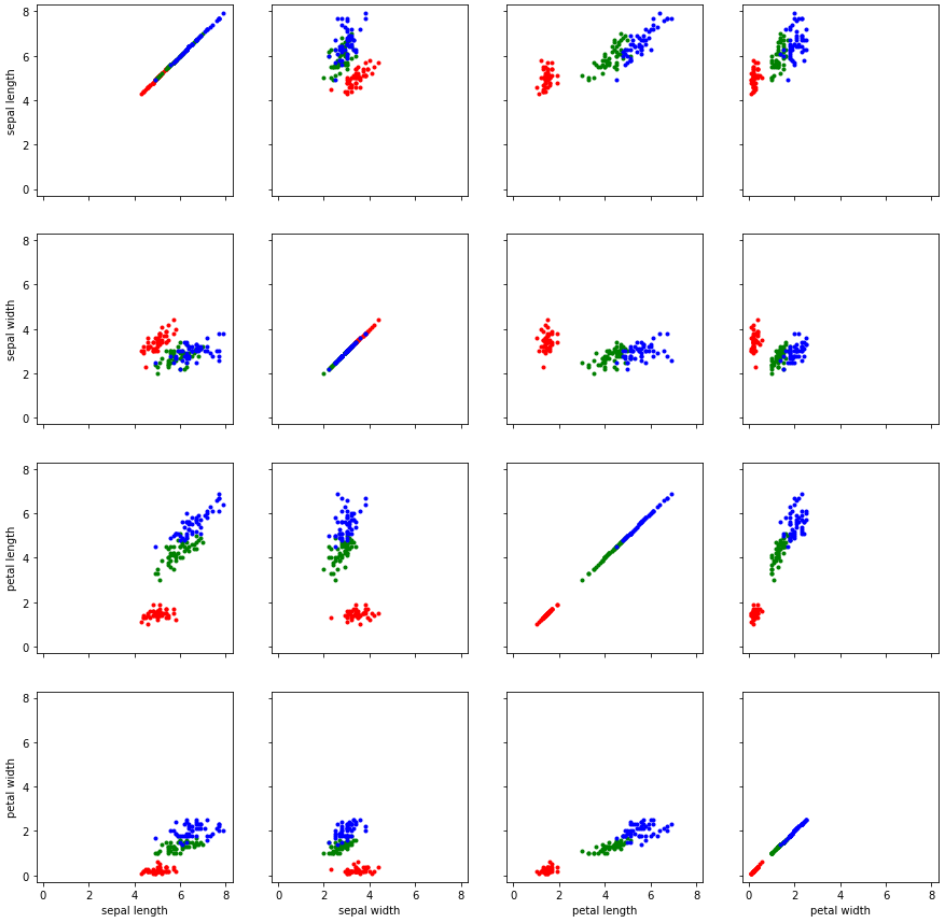
D. Visualisasi Cluster

Implementasi visualisasi cluster menggunakan kakas pyplot dari pustaka matplotlib. Pada visualisasi dengan color-coding pada titik data, warna merah digunakan pada nilai prediksi cluster 0, warna hijau digunakan pada nilai prediksi cluster 1, dan warna biru digunakan pada nilai prediksi cluster 2.

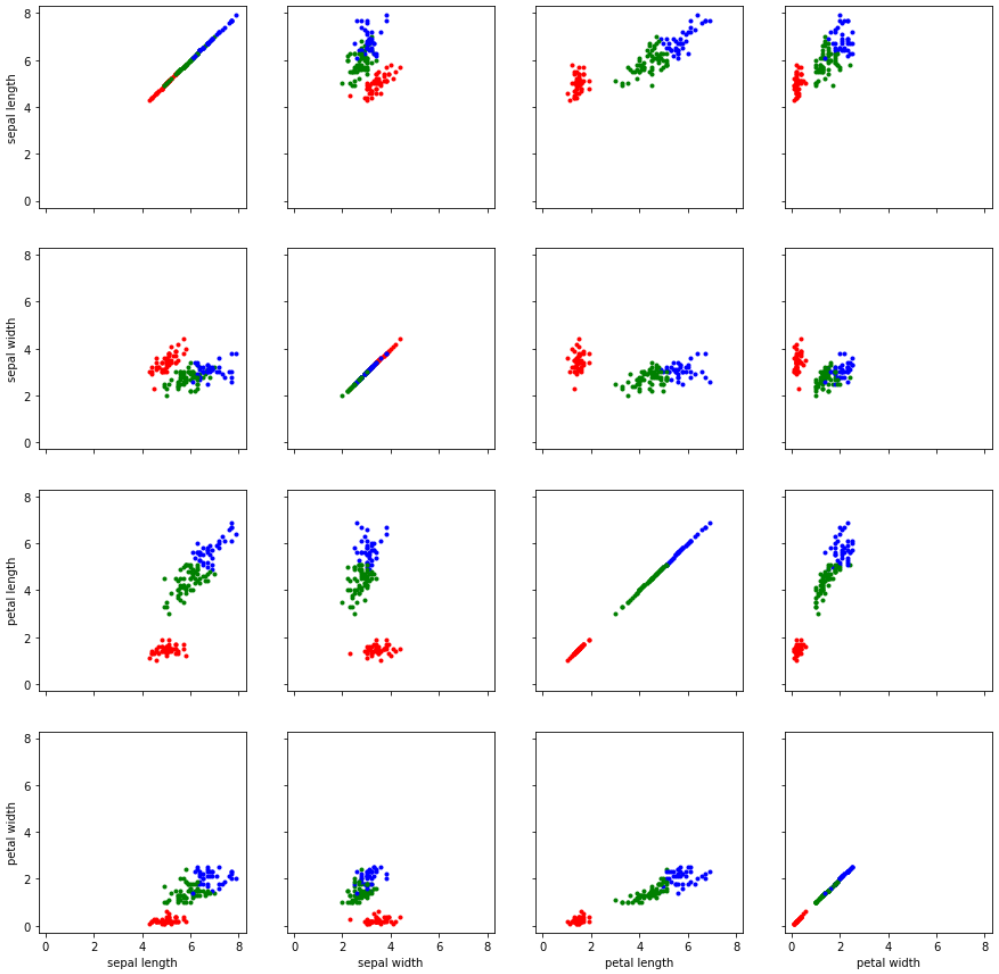
1. Plotting Data Latih
    a. Tanpa Color-coding Data Target Clustering



    b. Dengan Color-coding Data Target Clustering

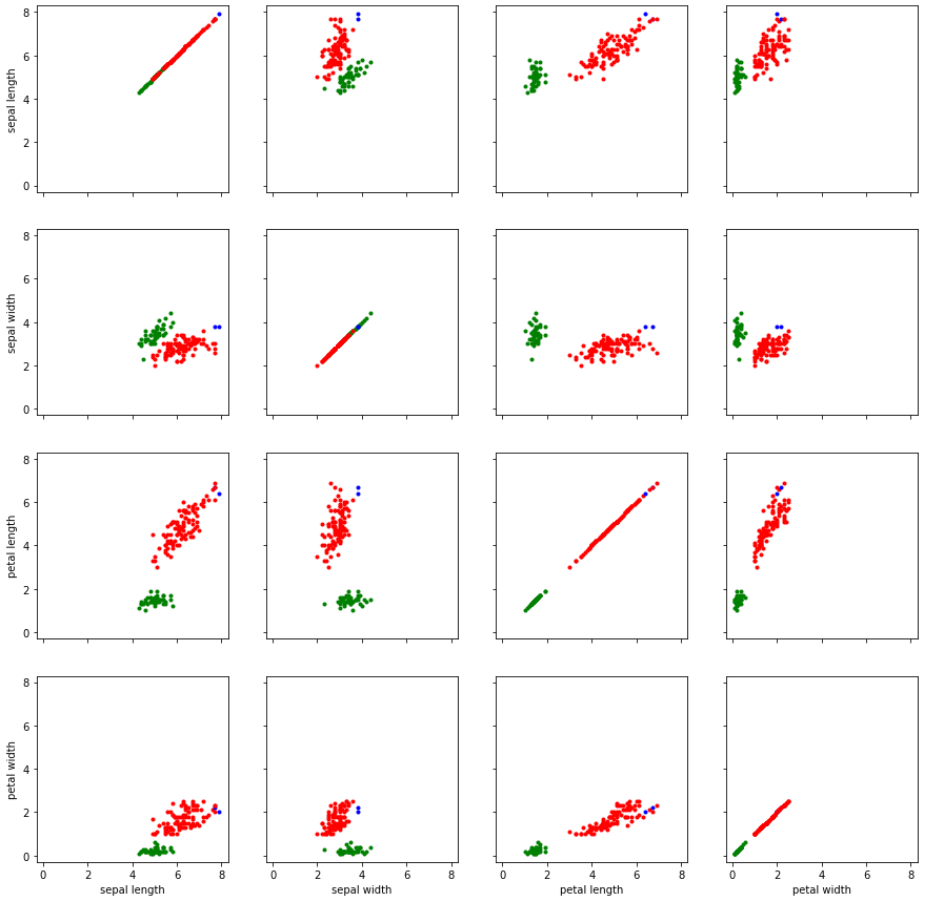


2. KMeans



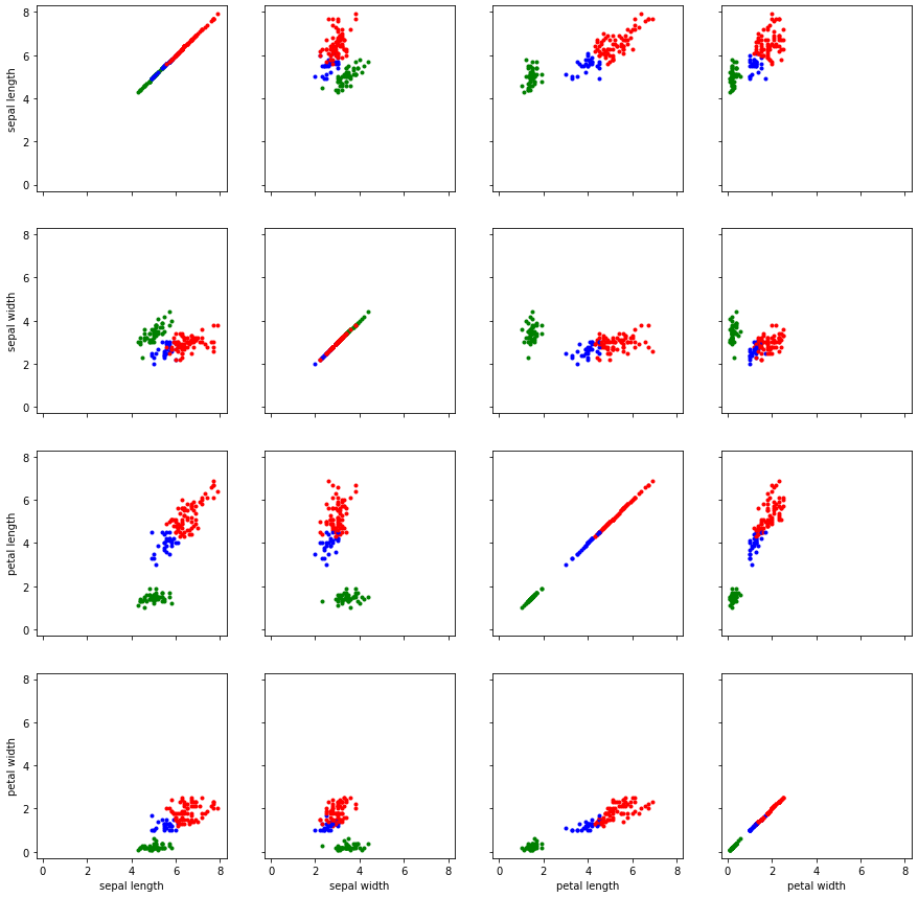
Dari visualisasi dataset dengan hasil klasterisasi dengan algoritma k-means, dibandingkan dengan data target (data_y), hasil klasterisasi dengan k-means mendekati data target pada dataset latih. Hanya saja, terdapat perbedaan dari segi urutan pembentukan cluster yang diakibatkan oleh inisialisasi centroid secara random. Perbedaan urutan pembentukan cluster ini terlihat dari cluster berwarna hijau dan merah yang tampak "tertukar" posisinya. Pada array hasil prediksi, titik berwarna merah adalah titik dengan nilai pred = 0 dan titik berwarna hijau adalah titik dengan nilai pred = 1.

3. Agglomerative
    a. Single Linkage



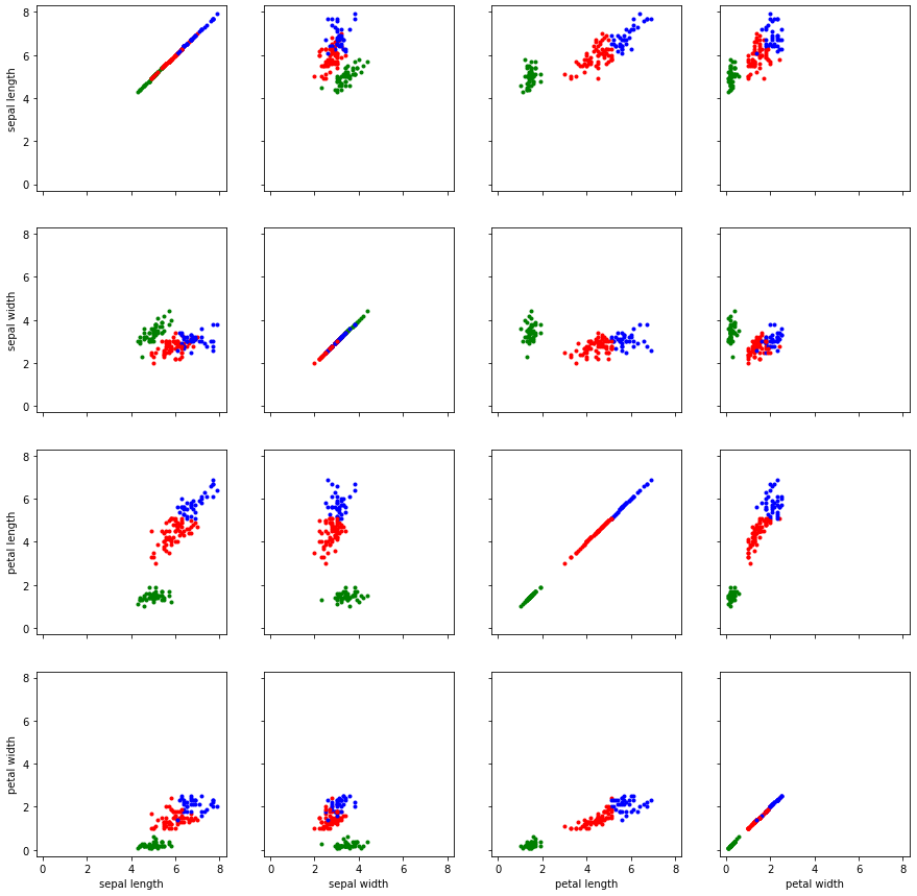
Dari visualisasi dataset dengan hasil klasterisasi dengan algoritma agglomerative - single linkage, dibandingkan dengan data target (data_y), hasil klasterisasi dengan agglomerative - single linkage relatif lebih berbeda dibandingkan dengan hasil klasterisasi dengan algoritma k-means. Hal ini terlihat dari kluster dengan titik-titik berwarna merah (pred = 0) yang sangat banyak dan titik-titik berwarna biru (pred = 2) yang sangat sedikit. Kluster dengan titik berwarna hijau relatif tidak mengalami perubahan jumlah yang signifikan dibandingkan dengan hasil algoritma k-means.

b. Complete Linkage



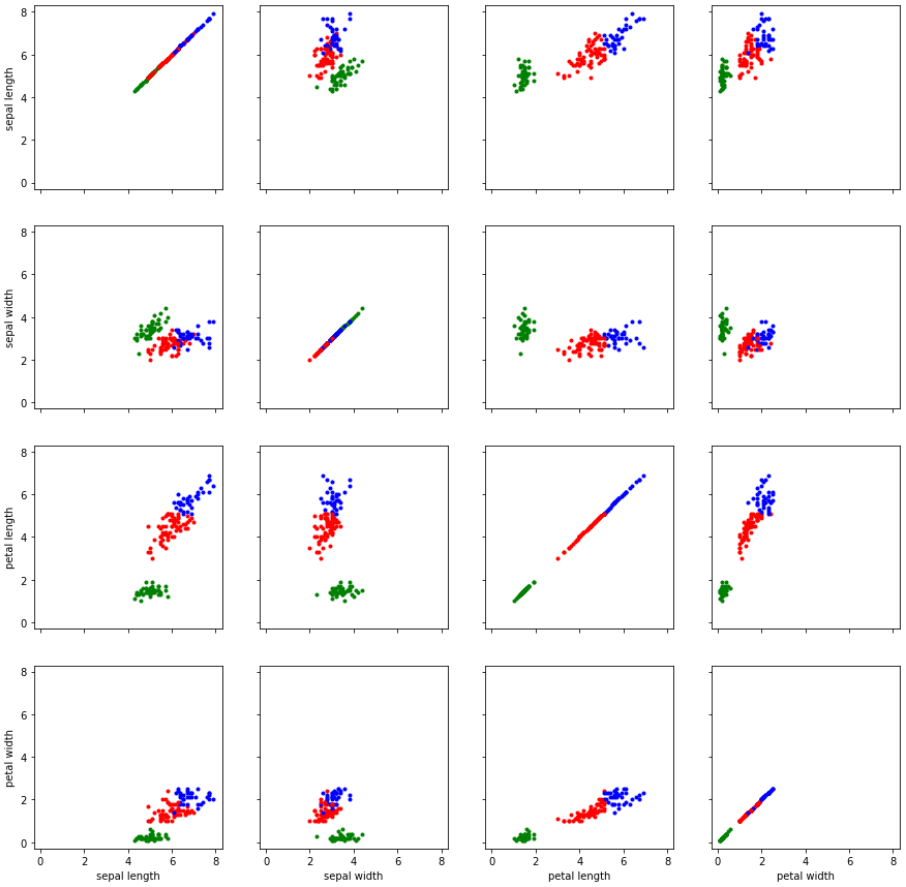
Dari visualisasi dataset dengan hasil klasterisasi dengan algoritma agglomerative - complete linkage, dibandingkan dengan data target (data_y), hasil klasterisasi dengan agglomerative - complete linkage memiliki perbedaan yang terletak pada lebih banyaknya titik berwarna biru dibandingkan titik berwarna hijau pada data target (yang menunjukkan group atau klaster yang analog). Hasil klasterisasi dengan agglomerative - complete linkage juga berbeda dengan agglomerative - single linkage, di mana jumlah anggota klaster-klaster pada complete linkage tidak setimpang pada single linkage.

c. Average Linkage



Dari visualisasi dataset dengan hasil klasterisasi dengan algoritma agglomerative - average linkage, dibandingkan dengan data target (data_y), hasil klasterisasi dengan agglomerative - average linkage memiliki perbedaan yang terletak pada lebih banyaknya titik berwarna merah dibandingkan titik berwarna hijau pada data target (yang menunjukkan group atau klaster yang analog). Hasil klasterisasi dengan agglomerative - average linkage juga berbeda dengan agglomerative - single linkage, di mana jumlah anggota klaster-klaster pada average linkage tidak setimpang pada single linkage. Walaupun begitu, hasil klasterisasi pada average linkage memiliki titik biru yang lebih sedikit dibandingkan titik merah (klaster yang analog) pada complete linkage.

d. Average Group Linkage



Terlihat bahwa klaster-klaster yang dihasilkan algoritma agglomerative - average group linkage identik dengan klaster-klaster pada algoritma agglomerative - group linkage. Penyebabnya adalah ukuran dataset yang kecil sehingga penentuan jarak antar cluster dengan menghitung rata-rata dari setiap kemungkinan jarak data antar cluster (pada average linkage) hasilnya sangat mendekati penentuan jarak dengan menghitung rerata dari masing-masing cluster (centroid) terlebih dahulu.

E. Pembagian Tugas

NIM	**Nama**	**Pembagian Tugas**
13517029	Reyhan Naufal Hakim	Visualisasi klaster untuk data latih, visualisasi hasil prediksi algoritma k-means, visualisasi hasil prediksi algoritma Agglomerative (single, complete, average, group average linkage); analisis hasil visualisasi klaster pada python notebook; Laporan bagian visualisasi cluster
13517035	Hilmi Naufal Yafie	MyAgglomerative, Evaluasi dengan Fowlke-Mallows dan Silhouette Coefficient

		algoritma Agglomerative, Laporan bagian implementasi MyAgglomerative.
13517095	Naufal Zhafran Latif	Implementasi algoritma KMeans, Laporan bagian implementasi algoritma kmeans, inisiasi struktur projek.
13517098	Anzaldi Sulaiman Oemar	Evaluasi Fowlke-Mallows dan Silhouette Coefficient algoritma K-Means, evaluasi Fowlke-Mallows dan Silhouette Coefficient algoritma Agglomerative, implementasi clustering iris dataset pada jupyter notebook, laporan bagian evaluasi.