K-MEANS CLUSTERING
MATA KULIAH PEMBELAJARAN MESIN

Disusun oleh:
Ainun Nisa
1301154527 / IF39-09

UNIVERSITAS TELKOM
BANDUNG
2018

A. Landasan Teori

"Algoritma K-Means adalah salah satu algoritma unsupervised learning yang paling sederhana yang dikenal dapat menyelesaikan permasalahan clustering dengan baik" (Mac Queen, 1967)[1]

Algoritma pengelompokan (*clustering*) semacam K-Means dikenal sebagai algoritma pembelajaran tidak terbimbing (*unsupervised learning*) karena tidak adanya target kelas untuk setiap data. Algoritma K-Means menjadi salah satu algoritma yang peling penting dalam bidang data mining. Memiliki kelebihan sebagai algoritma yang mudah diimplementasikan, relatif cepat ditinjau dari waktu komputasi dan telah digunakan secara luas untuk menyelesaikan berbagai persoalan komputasi.

Penting untuk diketahui bahwa, K yang dimaksudkan dalam K-Means adalah jumlah partisi, sehingga algoritma ini mengelompokkan setiap titik pada data X dalam salah satu partisi K. Semakin besar nilai K berdampak pada banyaknya cluster untuk mengelompokkan data X. Nilai K ini juga menjadi parameter yang dibutuhkan oleh algoritma K-Means. Tidak ada ketetapan mutlak bagaimana menentukan nilai K yang optimal. Biasanya, penentuan nilai K didasarkan atas informasi yang diketahui sebelumnya tentang seberapa banyak cluster data yang muncul pada data X. Cara lain yang paling naif dalam menentukan nilai K adalah melakukan percobaan dengan beberapa nilai K hingga ditemukan bentuk pengelompokan yang tepat. Namun dewasa ini, banyak peneliti telah mencoba untuk melakukan pencarian nilai K yang optimum pada K-Means menggunakan berbagai pendekatan meta-heuristik seperti particle swarm optimization (pso), ant colony optimization (aco) dan lain sebagainya.[2]

B. Analisis Masalah

Pada tugas kali ini, diberikan sujumlah buah data, dimana data tersebut dibagi menjadi data train dan data test. Dari data tadi, diperintahkan untuk membangun sebuah sistem *K-Means Clustering* untuk mengklasterisasi data yang diberikan.

C. Desain

Istilah yang lazim ditemukan dalam algoritma K-Means adalah **centroid**. Centroid dapat diartikan sebagai titik pusat cluster, banyaknya centroid juga bergantung dari banyaknya nilai K yang diberikan. Gambar 1 menunjukkan terdapat data set yang

dilakukan pengelompokan menjadi dua cluster. Maka, untuk mendapatkan dua cluster, nilai K yang diberikan adalah 2. Kemudian, dua cluster tersebut harus memiliki centroid sebagai titik pusat cluster. Penentuan centroid ini dapat dilakukan secara random atau mengambil beberapa titik data sebagai centroid. Data akan terkelompok berdasarkan centroid terdekat, sehingga terbentuklah suatu pengelompokan data sesuai jumlah K yang diberikan.

Secara garis besar, algoritma K-Means Clustering dijelaskan dalam 5 tahap berikut :

1. Inisialisasi, tentukan nilai K sebagai jumlah cluster. Jika perlu tetapkan ambang batas perubahan fungsi objektif (batas yang menentukan iterasi berhenti atau tidak) dan ambang batas perubahan posisi centroid.
2. Pilih K data dari data set X sebagai centroid.
3. Alokasikan semua data ke centroid terdekat dengan menghitung metrik jarak.
4. Hitung kembali centroid C berdasarkan data yang mengikuti cluster masing-masing.
5. Ulangi langkah 3 dan 4 hingga kondisi konvergen tercapai, yaitu (a) perubahan fungsi objektif sudah di bawah ambang batas atau (b) tidak ada data yang berpindah cluster atau (c) perubahan posisi centroid sudah berada di bawah ambang batas. [2]

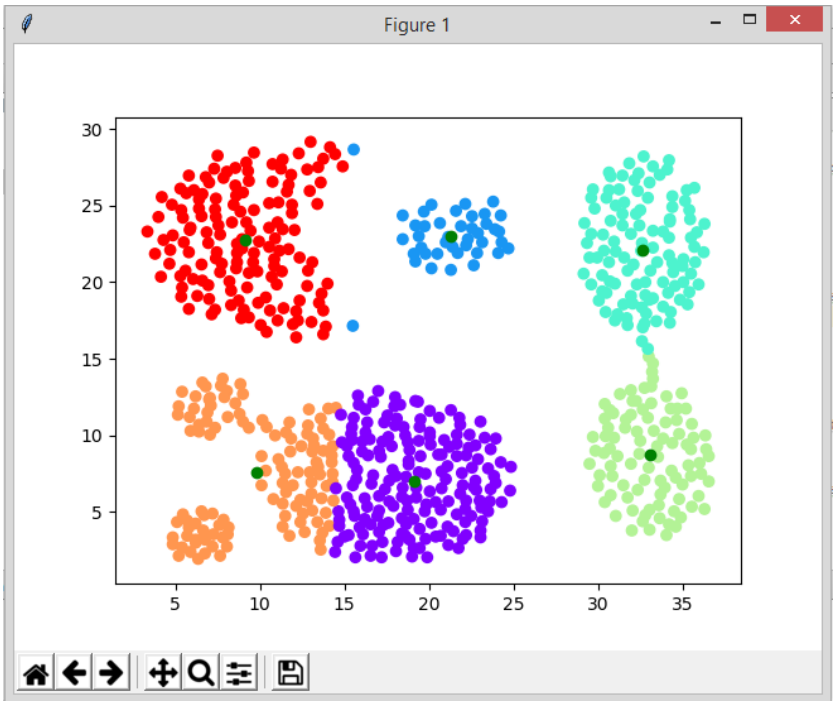
Disini penulis membangun sistem k-means clustering mengunakan *scikit learn Python* yang memudahkan dengan library yang disediakan. Awalnya melakukan load terhadap data train, lalu mengambil data dari masing-masing kolom/atribut yang diberi nama sesuai dengan yang ada dalam file dan melakukan plot terhadap data tersebut.

Tahap selanjutnya, dilanjutkan menggunakan pendekatan *scikit-learn* untuk menentukan centroidnya. Setelah itu barulah dilakukan load dan proses data untuk Testset yang diberikan. Visualisasinya, untuk masing-masing data dengan label yang berbeda, diberi warna berbeda dengan kodewarna rainbow, sedangkan centroid dibuat berbeda warna agar dapat dilihat titik centroidnya ada di sebelah mana.

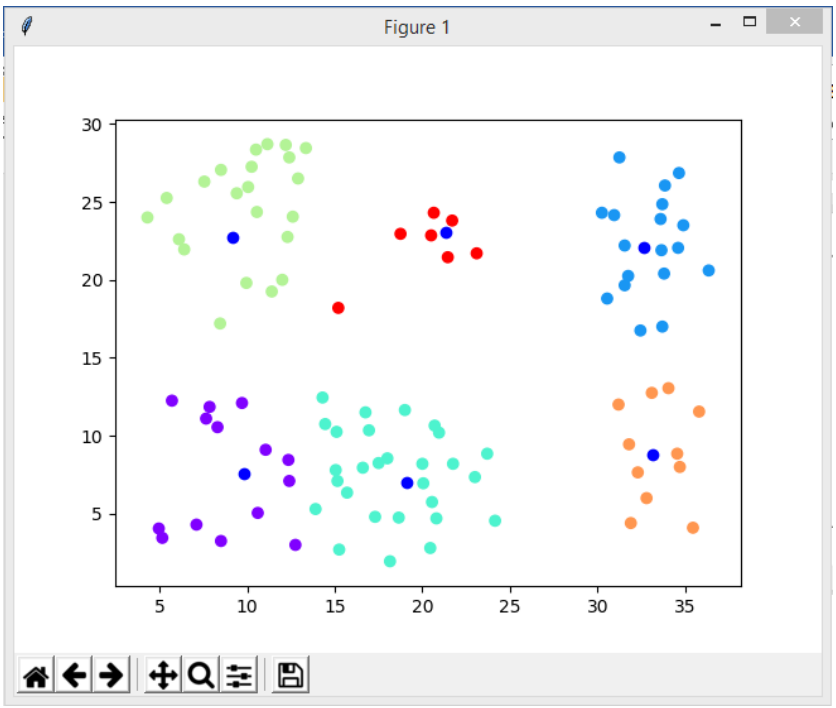
D. Hasil Program

Adapun hasil dari program yang coba penulis buat adalah sebagai berikut:

- Hasil Trainset



- Hasil Testset



Label dari masing-masing data atau hasil dari dataTest ada dalam file hasil.txt.

E. Evaluasi

Setelah memvisualisasikan data Train, penulis mencoba untuk memilih beberapa jumlah jumlah K, setelah dilakukan percobaan pada K=6, data Train terlihat cocok, namun ada beberapa data yang menjauh dari kumpulan data lain pada centroid yang sama. Tapi tetap dipilih K=6 karena jika dipilih K yang dianggap terlalu pas untuk Train, ditakutkan akan terjadi overfitting saat diimplementasikan pada data testnya.

F. Referensi

Fauzan, A. (2016, December 13). *Algoritma K-Means Clustering dan Contoh Perhitungan Untuk Data Numerik 2 Dimensi (bag. 1)*. Retrieved from Informatika Kita: http://www.charisfauzan.net/2016/12/algoritma-k-means-clustering-dan.html

Nango, D. N. (2012). PENERAPAN ALGORITMA K-MEANS. *Skripsi*, 2.

NK, M. (2017, October 1). *K-means Clustering in Python*. Retrieved from Mubaris's Blog: https://mubaris.com/2017/10/01/kmeans-clustering-in-python/

sklearn.cluster.KMeans. (n.d.). Retrieved from scikit learn: http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html