**TUGAS 2 K-MEANS CLUSTERING**



**Fero Resyanto**

**1301154318**

**IF-39-10**

**JURUSAN S1 TEKNIK INFORMATIKA**

**FAKULTAS INFORMATIKA**

**UNIVERSITAS TELKOM**

**BANDUNG**

**2018**

Daftar isi

[**Deskripsi Masalah** 3](#_Toc510699034)

[**Rancangan Metode / Design** 3](#_Toc510699035)

[**Evaluasi** 4](#_Toc510699036)

[**Referensi** 6](#_Toc510699037)

# **Deskripsi Masalah**

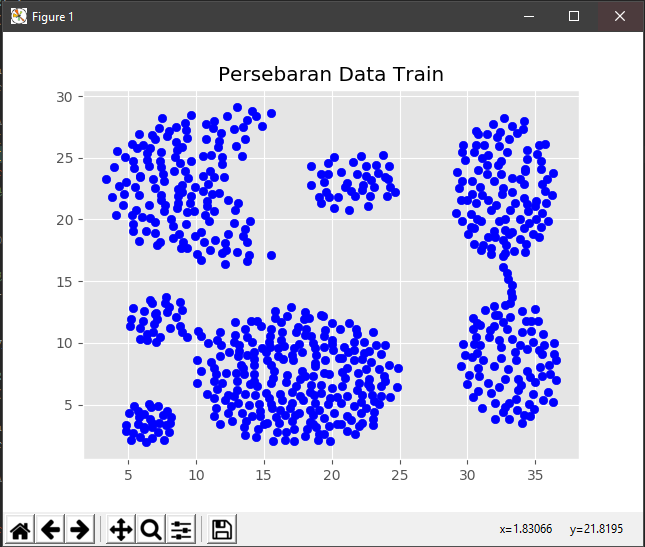
Pada kasus kali ini, kita menghadapi sebuah masalah yaitu bagaimana mengklusterkan data test yang diberikan. Untuk menyelesaikan masalah tersebut, kita dapat menggunakan algoritma *learning* k-means untuk mencari cluster masing-masing setiap data. Algoritma k-means itu sendiri merupakan salah satu metode clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk suatu cluster. Metode ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karateristik yang berbeda di kelompokan ke dalam cluster yang lain. Algoritma k-meams termasuk pemodelan *unsupervised* dan termasuk *non-deterministic* ( nilai akan berubah untuk setiap *running* program).

# **Rancangan Metode / Design**

Secara umum algoritma dasar dari K-Means Clustering adalah sebagai berikut :

1. Tentukan jumlah cluster.
2. Alokasikan data ke dalam cluster secara random.
3. Hitung centroid/rata-rata dari data yang ada di masing-masing cluster.
4. Alokasikan masing-masing data ke centroid/rata-rata terdekat.
5. Kembali ke step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan.

Pada kasus ini nilai K yang digunakan adalah 5, sehingga kita memiliki 5 cluster. Untuk menentukan nilai centroid yang akan digunakan pada data test, kita harus mencari nilai centroid optimal terlebih dahulu yang digunakan pada data train agar nilai centroid tersebut dapat digunakan pada data test.

Berikut ini adalah visualisasi persebaran data train menggunakan scatter plot :

Setelah itu, kita menentukan centroid awal dengan cara memilih dari data train secara random untuk dijadikan cetroid awal. Kemudian, untuk setiap data train, dihitung jaraknya untuk setiap centroid yang ada. Lalu pilih jarak terpendek yang mana centroid dengan jarak terpendek tersebut dijadikan sebagai cluster untuk data train tersebut. Setelah setiap data train memiliki cluster, hitung centroid atau rata-rata dari data train yang ada di masing-masing cluster.

Nilai yang didapat dari rata-rata tersebut digunakan sebagai nilai centroid baru pada iterasi berikutnya. Hal tersebut dilakukan sampai batas iterasi yang telah ditentukan sebelumnya. Sehingga diakhir iterasi, kita memiliki nilai centroid yang dapat digunakan pada data test nanti.

# **Evaluasi**

Berikut ini nilai centroid yang didapat pada saat melakukan *running* program :

Centroid 1 : 10.0 , 22,7

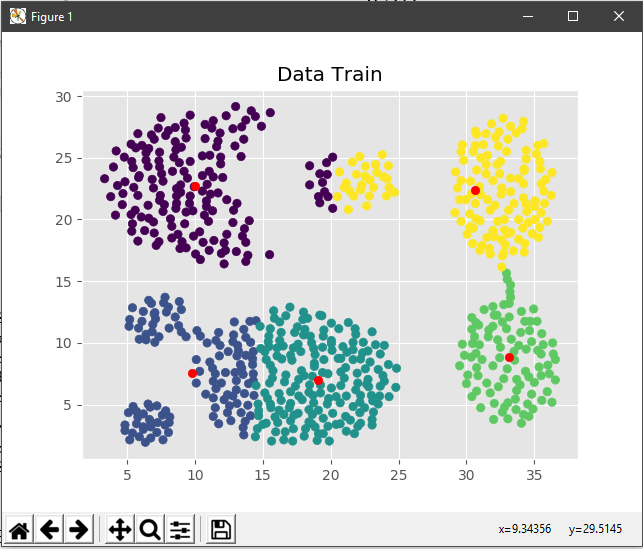
Centroid 2 : 9.8 , 7.5

Centroid 3 : 19.0 , 7.0

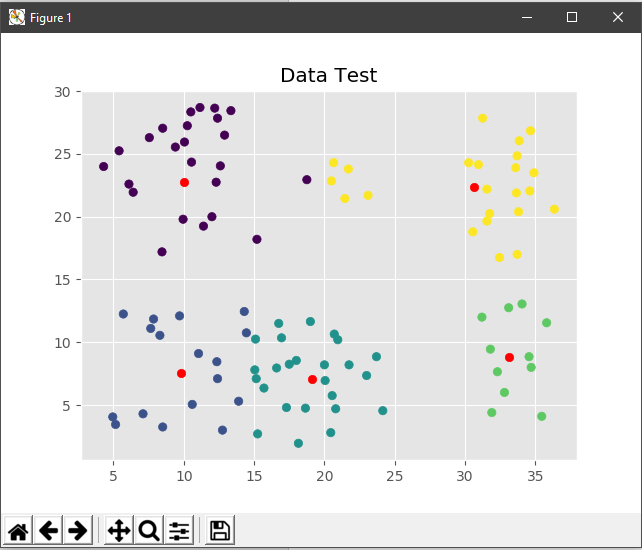
Centroid 4 : 33.1 , 8.8

Centroid 5 : 30.6 , 22.3

Sehingga hasil yang didapatkan adalah sebagai berikut :



Setelah mendapatkan nilai centroidnya, maka kita dapat gunakan pada data test yang kita miliki. Berikut hasil yang didaptkan :



Catatan :

* Warna merah menunjukan nilai centroid yang didapatkan.
* Warna ungu menunjukan data train/test dengan cluster = 1
* Warna biru menunjukan data train/test dengan cluster = 2
* Warna tosca menunjukan data train/test dengan cluster = 3
* Warna hijau muda menunjukan data train/test dengan cluster = 4
* Warna kuning menunjukan data train/test dengan cluster = 5
* Dikarenankan algoritma k-means *non-deterministic* maka, ketika melakukan *running* program hasilnya dapat berbeda dengan laporan ini.

Cluster hasil prediksi dapat dilihat pada file dengan nama : *‘hasil\_data\_test.txt’* (lakukan running program terlebih dahulu).

# **Referensi**

* <http://sharewy.blogspot.co.id/2013/04/algoritma-k-means-clustering.html>
* <https://informatikalogi.com/algoritma-k-means-clustering/>
* <https://www.youtube.com/watch?v=wt-X61BnUCA&t=301s>