



**Prosiding
Seminar Nasional Komputasi 2012**

**Menjawab Tantangan Dunia
dengan
Ilmu Komputasi**

Bandung, 22 september 2012

**Prodi Ilmu Komputasi
Institut Teknologi Telkom
snakom.it Telkom.ac.id**

**Penerbit
ISBN : 978-602-18892-0-6**



**INSTITUT TEKNOLOGI
TELKOM**

**PERFORMANCE DISTANCE SPACE
MANHATTAN(CITYBLOCK) DENGAN EUCLIDEAN
PADA ALGORITMA K-MEANS CLUSTERING
STUDI KASUS : DATA BALITA DI WILAYAH KEC.MLATI SLEMAN**

Sri Redjeki¹, Andrias Pamungkas²

^{1,2}Jurusan Teknik Informatika, STMIK AKAKOM Yogyakarta
redjeki@akakom.ac.id, dekomp@yahoo.com.

Abstrak

Teknik Clustering merupakan salah satu metode Data Mining yang bersifat tanpa arahan (*unsupervised learning*). Tujuan utama dari metode cluster adalah pengelompokan sejumlah data/objek ke dalam kelompok – kelompok data sehingga dalam setiap kelompok akan berisi data yang semirip mungkin. Salah satu algoritma yang paling sederhana clustering adalah K-Means. Metode ini mempartisi data ke dalam cluster sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dengan cara menghitung jarak terdekat data dengan titik pusat data/*centroid* yaitu dengan rumus *distance space*.

Beberapa *distance space* yang ada antara lain Manhattan(CityBlock), Euclidean dan Minkowski. Masing-masing *distance space* tersebut mempunyai kemampuan yang cukup berbeda sehingga perlu dilakukan perbandingan *performance* jenis-jenis *distance space* yang ada. Pada penelitian ini akan dibandingkan *performances* dari tiga *distance space* yaitu antara Manhattan(City Block) dengan Euclidean pada Clustering K-Means dengan objek data status gizi balita di Kecamatan Mlati Sleman.

Hasil akhir percobaan didapatkan bahwa pola clustering dari kedua rumus tersebut adalah berbeda, perbedaan ini terlihat pada hasil jumlah iterasi, serta perbandingan waktu. Manhattan(CityBlock) cenderung lebih cepat dibandingkan dengan rumus Euclidean namun dari beberapa percobaan hasil proses clustering Euclidean lebih baik dari Manhattan(CityBlock).

Kata Kunci : Clustering, Euclidian, Gizi Balita, K-Means, Manhattan (City Block), Performance.

Abstract

Clustering technique is one of data mining method, it character is without direction/unsupervised learning. The main objective of the cluster is for grouping several methods of data objects which each group will contain the similar / resembled data. One of the simplest Clustering algorithm is the K-Means. This method partitioning data into clusters so the data with the same characteristic can be grouped into the same cluster by calculating between the shortest distance of data and the data center/centroid with *distance space* formula.

The existing *distance space* formulas are Manhattan(CityBlock), Euclidean and Minkowski. Each *distance space* tend to have different ability so it is necessary to compare their performance. This research will compare two of three *distance space*, it is between Manhattan(City Block) and the Euclidean K-Means Clustering with nutritional data of children in the District Mlati Sleman as data objects.

The End of the experiment results obtained that the clustering pattern of the two formulas are different, this difference was seen in the number of iterations, as well as the comparison of time. Manhattan(CityBlock) tend to be faster than the Euclidean formula, but some experimental results of Euclidean clustering process is better than Manhattan(CityBlock).

Keywords: Clustering, Euclidian, Nutrition, K-Means, Manhattan (City Block), Performance.

1. Pendahuluan

Teknik Clustering merupakan salah satu metode Data Mining yang bersifat tanpa arahan (*unsupervised learning*) dimana tidak perlu ada *flow chart* (*flowing*) pada metode tersebut. Tujuan utama dari metode cluster adalah pengelompokan sejumlah data/objek ke dalam kelompok – kelompok data sehingga dalam setiap kelompok akan berisi data yang semirip mungkin (Dodi Samson, 2007). Ada dua pendekatan dalam clustering yang sering

digunakan dalam proses pengelompokan data yaitu *hierarchical* (*hirarki*) data clustering dan *partitional* (*non hirarki*) data clustering. K-Means merupakan salah satu metode data clustering non hirarki yang paling sederhana.

Menurut Yudi Agusta (2007) 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 karakteristik yang berbeda

dikelompokkan ke dalam kelompok yang lain dengan cara menghitung jarak terdekat data dengan titik pusat data/centroid yaitu dengan rumus *distance space*.

Beberapa *distance space* telah diimplementasikan dalam menghitung jarak (*distance*) antara data dan *centroid* termasuk di antaranya adalah L1 (*Manhattan / City Block distance space*), L2 (*Euclidean distance space*), dan Lp (*Minkowski distance space*). Tetapi secara umum *distance space* yang sering digunakan adalah *Manhattan/City Block* dan *Euclidean*. *Euclidean* sering digunakan karena penghitungan jarak dalam *distance space* ini merupakan jarak terpendek yang bisa didapatkan antara dua titik yang dipertanyakan, sedangkan *Manhattan/City Block* sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus seperti keberadaan *outliers* dengan lebih baik (Yudi Agusta, 2007).

Pada penelitian ini, akan dibandingkan dua *distance space* yaitu antara *Manhattan/City Block* dan *Euclidean* dengan objek penelitian yaitu data anak balita usia dibawah lima tahun. Dari dua rumus tersebut akan dipelajari hasil pengelompokan data yang terbentuk.

Adapun tujuan dari dibuatnya penelitian ini yaitu mengetahui *performances* hasil clustering data menggunakan metode K-Means dengan rumus *distance space* antara *Manhattan/City Block* dan *Euclidean*, serta menggali informasi data dari objek penelitian data status gizi balita di wilayah Mlati Sleman. Sehingga pada hasil akhir *cluster* diharapkan dapat dipelajari pola pengelompokan mana yang lebih baik untuk masing-masing rumus *distance space*. Identifikasi masalah yang ada pada penelitian ini antara lain: melakukan perbandingan antara dua *distance space* pada K-Means yaitu *Manhattan/City Block* dengan *Euclidean*, jumlah *cluster* yang digunakan pada penelitian sebanyak 4 *cluster* yang menggambarkan status gizi balita, data yang digunakan untuk adalah data balita di wilayah Puskesmas Kecamatan Mlati Sleman Yogyakarta.

Metode penelitian dilakukan dengan cara antara lain: survei data mengenai status gizi balita yang ada di Kabupaten Sleman terutama Kecamatan Mlati, melakukan *preprocessing* data sebelum dilakukan *training*, melakukan desain sistem berbasis K-Means, melakukan *training* dan *testing* data objek penelitian untuk dua jenis *distance space* serta melakukan perbandingan hasil clustering dari 2 jenis *distance space*.

2. Pembahasan Penelitian

2.1 K-Means Clustering

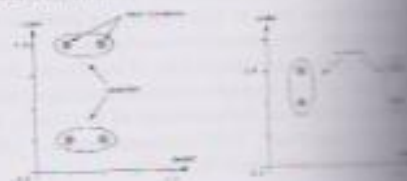
Teknik Clustering merupakan salah satu metode Data Mining yang bersifat tanpa arahan (*unsupervised*). Data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster*

yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Menurut Yudi Agusta (2007), tujuan utama dari metode *cluster* adalah pengelompokan sejumlah data/objek ke dalam kelompok - kelompok data sehingga dalam setiap kelompok akan berisi data yang semirip mungkin. Ada dua jenis data clustering yang sering digunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) dan *clustering* dan *non-hierarchical* (non hirarki) dan *clustering* atau yang disebut *partitioning*.

Dalam clustering hirarki, akan dimulai dengan membuat *m* *cluster*, dimana setiap *cluster* beranggotakan satu objek dan berakhir dengan satu *cluster* dimana anggotanya adalah *m* objek dan berakhir dengan satu *cluster* dimana anggotanya adalah *m* objek. Pada setiap tahap dalam prosedurnya, satu *cluster* digabung dengan *cluster* yang lain. Sedangkan dalam *partitioning/non hirarki*, data dikelompokkan ke objek - objek ke dalam *k* *cluster*.

K-Means merupakan salah satu metode dan clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster* atau kelompok. Metode ini mempartisi data ke dalam kelompok - kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu *cluster* yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain dengan cara menghitung jarak terdekat data dengan titik pusat data/centroid yaitu dengan rumus *distance space*.

Distance space atau jarak objek dengan titik pusat menjadi komponen yang sangat penting dalam algoritma clustering. Jarak terdekat akan menjadi penentu objek untuk masuk ke dalam *cluster* yang ada. Ilustrasi *distance space* dapat kita lihat pada gambar dibawah.



Gambar 1. Ilustrasi Distance Space

Yudi agusta (2007) dalam makalahnya mengatakan, beberapa rumus *distance space* yang diimplementasikan dalam menghitung jarak (*distance*) antara data dan *centroid* termasuk di antaranya L1 (*Manhattan/City Block distance space*), L2 (*Euclidean distance space*), dan Lp (*Minkowski distance space*). Jarak antara x_1 dan x_2 pada *Manhattan/City Block distance space* dihitung dengan menggunakan rumus sebagai berikut:

$$D_{L_1}(x_2, x_1) = \|x_2 - x_1\|_1 = \sum_{j=1}^p |x_{2j} - x_{1j}| \quad (1)$$

Sedangkan untuk L_2 (Euclidean) distance space, jarak antara dua titik dihitung menggunakan rumus sebagai berikut:

$$D_{L_2}(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^p (x_{2j} - x_{1j})^2} \quad (2)$$

dimana:

p : Dimensi data dan $||$: Nilai absolute

Budi Santoso (2007) menyatakan selain mean atau rata-rata sebagai prosedur pengelompokan, dikenal juga ukuran pemusatan yang lain seperti median yang bisa digunakan. Untuk kasus – kasus tertentu pemakaian median sebagai alternatif dari mean memberikan hasil yang lebih baik. Seperti kita ketahui median tidak sensitif terhadap data outlier, yaitu data yang terletak jauh dari sekumpulan data.

Clustering menggunakan metode K-Means secara umum dilakukan dengan algoritma dasar sebagai berikut:

1. Tentukan jumlah cluster
2. Inisialisasi titik pusat (*centroid*) data ke dalam cluster secara random
3. Hitung jarak titik tengah dengan data menggunakan rumus distance space
4. Alokasikan masing-masing data ke *centroid* berdasarkan hasil distance space
5. Hitung kembali titik pusat cluster dengan keanggotaan data cluster yang sekarang. Titik pusat adalah rata-rata dari semua data dalam cluster tertentu
6. Apabila masih ada data yang berpindah cluster maka kembali ke Step 3, jika tidak ada maka proses clustering selesai

Hasil cluster dengan metode k-means sangat bergantung pada nilai pusat cluster yang diberikan. Perubahan nilai awal yang berbeda bisa menghasilkan hasil cluster yang berbeda. Adapun tujuan dari data clustering ini adalah untuk meminimalisasikan *objective function* yang disebut dalam proses clustering, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu cluster dan memaksimalkan variasi antar cluster.

3.2 Model Penelitian

Clustering menggunakan K-Means mempunyai kelemahan dimana iterasi pada algoritma ini akan berhenti apabila sudah tidak ada perubahan nilai suatu cluster maupun data yang ada pada cluster akan berubah. Selama proses iterasi ini, jumlah data

yang ada pada masing-masing cluster akan terus berubah sampai mendapatkan jarak terbaiknya.

Penelitian ini mempunyai model penelitian yang di gambarkan menggunakan blok diagram yang ada pada gambar 2 (Sri, 2010)



Gambar 2. Blok Diagram K-Means Clustering

Pada gambar 2 proses iterasi pada algoritma K-Means terlihat pada blok yang ada pada kotak K-Means, selama nilai ini tengah tidak ada perubahan maka proses iterasi akan berlanjut terus dan akan berhenti apabila sudah tidak ada perubahan nilai titik tengah. Sistem Clustering yang dibangun pada penelitian ini dapat dilihat pada diagram aktivitas pada gambar 3.

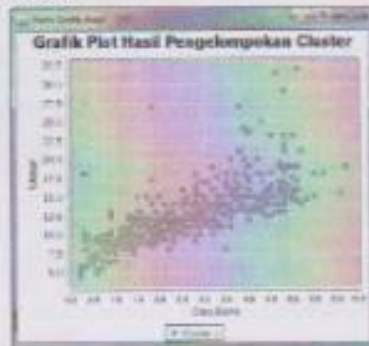


Gambar 3. Diagram Aktivitas Sistem

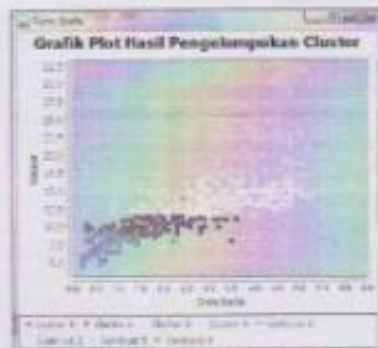
Pada gambar diatas menunjukkan diagram aktivitas proses clustering data, aktivitas akan dimulai dengan user menentukan jumlah cluster yang diinginkan. Sistem akan merandom nilai titik pusat data sesuai dengan jumlah cluster. Kemudian user memilih rumus distance space yang digunakan, rumus tersebut akan digunakan oleh sistem untuk menghitung jarak titik pusat dengan data dan mengoklasikan data berdasar jarak kedekatan data. Aktivitas selanjutnya adalah menghitung kembali titik pusat data dengan rumus mean atau rata – rata sesuai dengan keanggotaan cluster. Sistem akan mengecek apakah data masih ada yang berpindah cluster, jika ya maka proses kembali ke penghitungan jarak data dengan titik pusat, jika tidak maka proses akan selesai.

2.3 Analisis Hasil Penelitian

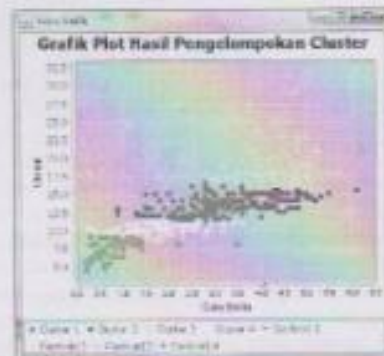
Pengujian aplikasi *clustering* menggunakan K-Means dilakukan pada data status gizi balita yang mewakili Kecamatan Mlati. Data balita yang dipilih sebanyak 500 data untuk membandingkan waktu dengan jumlah *cluster* sebanyak 4. Plot data obyek penelitian terlihat gambar 4, setelah dilakukan *clustering* menggunakan 2 *distance space* maka terlihat perbedaan hasil yang nampak pada gambar 5 dan gambar 6.



Gambar 4. Picting Data Awal
Perbedaan hasil *clustering distance space Euclidean* dan *Manhattan (CityBlock)*



Gambar 5. Hasil Grafik *Euclidean*



Gambar 6. Hasil Grafik *Manhattan (CityBlock)*

Dari hasil gambar 5 dan gambar 6 menunjukkan perbedaan hasil *clustering* yaitu hasil cluster 1 terdapat 72 data, cluster 2 terdapat 153 data, cluster 3 terdapat 65 data dan cluster 4 sebanyak 210 data, sedangkan untuk hasil *clustering* menggunakan *distance space Manhattan (CityBlock)* yaitu hasil cluster 1 terdapat 76 data, cluster 2 terdapat 203 data, cluster 3 terdapat 75 data dan cluster 4 terdapat 146 data. Perbedaan *performance* antara kedua *distance space* terlihat pada tabel 1.

Tabel 1. Perbandingan Waktu dan Iterasi

Periode iter.	Titik awal random	Iterasi	Jarak Euclidean		Manhattan (CityBlock)	
			Lama waktu (dalam menit)	Sangat Lama Waktu (dalam menit)	Lama waktu (dalam menit)	Sangat Lama Waktu (dalam menit)
1	40 = 2,1 41 = 2,4 42 = 2,1 43 = 2,4	75	402 menit	7	300 menit	
2	44 = 2,7 45 = 2,8 46 = 2,4 47 = 2,8 48 = 2,4 49 = 2,8	9	402 menit	8	370 menit	
3	49 = 2,8 50 = 2,8 51 = 2,8 52 = 2,8 53 = 2,8 54 = 2,8	9	402 menit	8	300 menit	

Diketahui bahwa perbandingan dari kedua rumus berdasarkan waktu akan bergantung kepada jumlah iterasi. Dilihat dari tabel diatas, rumus *manhattan (CityBlock)* lebih sedikit iterasinya sehingga lebih cepat beberapa milidetik dari rumus *Euclidean*.

Untuk mengetahui keakuratan hasil *clustering* dengan status gizi balita maka dilakukan percobaan *clustering* untuk 30 data. Hasil analisis sebagai berikut:

Tabel 2. Data Awal Status Gizi Balita

No	Umur	Berat	Gizi Lebih (GZL)	Gizi Rendah (GR)	Gizi Buruk (GB)	Gizi Tumbuh (GT)
1	2,2	6,4		✓		
2	2,2	11			✓	
3	2,2	11		✓		
4	2,2	11		✓		
5	2,2	10		✓		
6	2,2	18		✓		
7	2,2	7				✓
8	2,2	18	✓			
9	2,2	24,5		✓		
10	2,2	10			✓	
11	2,2	2,4				✓
12	2,2	8				✓
13	2,2	12			✓	
14	2,2	8,2	✓			
15	2,2	8		✓		
16	2,2	18,5	✓			
17	2,2	8,8			✓	
18	4,8	7,2	✓			
19	2,7	8,2		✓		
20	2,2	9			✓	
21	3,8	25	✓			
22	2,8	8,1		✓		
23	2,2	8		✓		
24	2,2	13,7		✓		
25	4,2	18		✓		
26	0,7	8		✓		
27	2,8	3				✓
28	2,2	18	✓			
29	2,2	5				✓
30	0,7	10		✓		

Hasil pengujian *Clustering* untuk 30 data pada *Euclidean* dan *Manhattan (City Block)* terlihat pada tabel 3 dan tabel 4.

Tabel 3. Hasil Akhir Euclidean

No	Dim	Beas	C1	C2	C3	C4	No	Dim	Beas	C1	C2	C3	C4
1	1.2	4.4				Y	18	2.0	3.1	Y			
2	1.3	11				Y	19	2.0	3.3				Y
3	2.5	11				Y	20	1.8	11				Y
4	1.4	11				Y	21	0.7	3.1				Y
5	1.8	11				Y	22	1.8	11				Y
6	1.2	11				Y	23	1.8	11				Y
7	1.8	7				Y	24	1.4	11.1				Y
8	1.8	11				Y	25	1.4	7				Y
9	1.7	11.3				Y	26	2.8	11.1				Y
10	1.1	11				Y	27	1.2	11				Y
11	1.3	4.3				Y	28	1.7	7				Y
12	1.4	4				Y	29	2.3	3				Y
13	1.3	11				Y	30	1.8	11				Y
14	1.2	11.1				Y							
15	1.3	7				Y							

Tabel 4. Hasil Akhir Manhattan (CityBlock)

No	Dim	Beas	C1	C2	C3	C4	No	Dim	Beas	C1	C2	C3	C4
1	1.2	4.4				Y	18	2.0	3.1				Y
2	1.3	11				Y	19	2.0	3.3				Y
3	2.5	11				Y	20	1.8	11				Y
4	1.4	11				Y	21	0.7	3.1				Y
5	1.8	11				Y	22	1.8	11				Y
6	1.2	11				Y	23	1.8	11				Y
7	1.8	7				Y	24	1.4	11.1				Y
8	1.8	11				Y	25	1.4	7				Y
9	1.7	11.3				Y	26	2.8	11.1				Y
10	1.1	11				Y	27	1.2	11				Y
11	1.3	4.3				Y	28	1.7	7				Y
12	1.4	4				Y	29	2.3	3				Y
13	1.3	11				Y	30	1.8	11				Y
14	1.2	11.1				Y							
15	1.3	7				Y							

Prosentase hasil dari Euclidean untuk tiap cluster yaitu dapat mengenali pada cluster 1 sebesar 44.4%, mengenali pada cluster 2 sebesar 100%, mengenali pada cluster 3 sebesar 53.84%, mengenali pada cluster 4 sebesar 42.85%. Untuk hasil dari Manhattan pada tiap cluster yaitu dapat mengenali pada cluster 1 sebesar 66.7%, mengenali pada cluster 2 sebesar 37.5%, mengenali pada cluster 3 sebesar 44.4% dan mengenali pada cluster 4 sebesar 50%.

Dapat kita lihat, bahwa dua rumus *distance space* antara *Manhattan(CityBlock)* dan *Euclidean* menghasilkan pola data yang berbeda untuk jumlah data sebanyak 30 dan jumlah kelompok sebanyak 4 cluster data. Dilihat dari hasil prosentase keakuratan data dengan tabel sama gizi, *Euclidean* cenderung lebih teliti dalam memastikan keanggotaan sebuah data ke dalam *cluster* walaupun jumlah iterasinya lebih banyak. Beberapa hal yang mempengaruhi hasil *cluster* terletak pada titik *centroid* awal yang berupa random.

Hasil *cluster* menunjukkan pola kecenderungan untuk tingkatan data yang tergantung

dari titik awal *centroid*. Tiap *cluster* akan mengikuti titik awal *centroid* sehingga tidak dapat dipastikan kelompok data akan masuk ke *cluster*. Oleh karena itu *clustering* data kurang cocok untuk menetapkan status gizi pada suatu wilayah. Algoritma *k-means* merupakan kategori algoritma pengelompokan data *unsupervised learning*.

3. Kesimpulan dan Saran

3.1. Kesimpulan

Dari hasil penjelasan pembahasan di atas maka dapat disimpulkan beberapa hal, antara lain:

- Hasil pengelompokan data oleh kedua rumus *distance space* *Manhattan(CityBlock)* dengan *Euclidean* adalah berbeda. Dari pola pengelompokan *Performance Euclidean* cenderung lebih teliti dalam memastikan keanggotaan sebuah data kedalam *cluster*. Pada percobaan pada 30 data balita *Euclidean* dapat mengenali status gizi hingga 60.27% sedangkan *Manhattan(CityBlock)* 48.4%.
- Waktu proses untuk kedua rumus tersebut selisih tidak jauh berbeda, karena akan bergantung dari jumlah iterasinya. Dari beberapa percobaan dengan 500 data balita, disimpulkan rumus *Manhattan(CityBlock)* memiliki jumlah iterasi yang lebih sedikit sehingga tukuran waktu lebih cepat beberapa mihdetik dalam mengelompokan data dari rumus *Euclidean*.
- Hasil dari random titik *centroid* awal akan sangat mempengaruhi hasil akhir *clustering*.

3.2. Saran

Dalam pengembangan penelitian *clustering K-Means* agar menjadi lebih baik lagi, penulis menyarankan beberapa hal:

- Membandingkan ketiga jenis *distance space* yang ada pada *K-Means* secara bersama-sama sehingga dapat diketahui performa masing-masing *distance space* yang ada.
- Menggunakan metode *Fuzzy C-Means*.
- Membandingkan *clustering K-Means* dengan *clustering* menggunakan *Neural Network*.

Daftar Pustaka:

- Bezdek, J. C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.
- Budi Santosa. 2007. *Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu, Yogyakarta.
- Miyamoto, S. and Agusta, Y. 1995. Algorithms for L1 and Lp Fuzzy C-Means and Their Convergence, in C. Hayashi, N. Oshimi, K. Yajima, Y. Tanaka, H. H. Bock and Y. Baba

- (eds), *Data Science, Classification, and Related Methods*, Springer-Verlag, Tokyo, Japan, pp. 295-302.
4. Sri Redjeki, dkk. 2010 "Clustering Terhadap Indeks Prestasi Mahasiswa Smpk Akademi Menggunakan K-Means".
 5. Yudi Agusta. 2007. *K-Means. Penerapan, Permasalahan Dan Metode Terkait* Jurnal Sistem dan Informatika 2007, Vol. 3.
 6. http://home.dei.palini.it/matteucci/Clustering/tutorial_html/kmeans.html