

Perbandingan Distance Space Manhattan Dengan Euclidean Pada K-Means Clustering Dalam Menentukan Promosi

Rahma Isparani Fajriah¹, Herlan Sutisna², Bambang Kelana Simpony³

^{1,3}Universitas Bina Sarana Informatika, ²Universitas BSI

Email: ¹rahmaisparani300@gmail.com, ²herlan.her@bsi.ac.id, ³bambang.bky@bsi.ac.id

Abstrak

Dinamika pola pendidikan yang begitu cepat dan silih berganti menjadikan persaingan antar sekolah semakin ketat dan atraktif, Dengan adanya persaingan ini strategi pemasaran yang tepat untuk lembaga pendidikan mutlak diperlukan, tak terkecuali untuk SMK Muhammadiyah Cimanggu. Untuk memperoleh strategi pemasaran yang efektif, maka penulisan ini dilakukan dengan cara mengolah data *history* dengan tujuan mengetahui pola, sehingga dapat mengambil informasi yang tersembunyi dari data-data tersebut. Metode pengolahan data seperti ini sering disebut sebagai data mining. K-Means merupakan salah satu metode data *clustering* non hirarki yang sederhana. Pada penulisan ini, membandingkan dua *distance space* yaitu antara *Manhattan (City Block)* dan *Euclidean*. Dari hasil penulisan dapat dilihat bahwa jumlah kluster terbaik pada data siswa SMK Muhammadiyah Cimanggu yaitu 3 kluster, hal ini dapat dilihat dari hasil selisih nilai *Sum Square Error* (SSE), nilai selisih terbesar pada metode *Euclidean distance space* yaitu sebesar 5.882 berada dititik kluster 3 dan nilai selisih terbesar pada metode *Manhattan distance space* yaitu sebesar 5.758 berada dititik kluster 3, Hal ini menunjukkan bahwa jumlah *cluster* optimal untuk data siswa SMK Muhammadiyah Cimanggu yaitu sebanyak tiga *cluster*. dilihat dari Average within *Centroid distance*, *Euclidean* lebih kecil dibandingkan dengan manhattaan yaitu $15.115 < 15.398$, oleh karna itu *Distance measure* paling optimal untuk digunakan dalam kasus pengclusteringan data siswa SMK Muhammadiyah Cimanggu adalah *Euclidean distance*.

Kata kunci: k means, *manhattan*, *euclidean*, strategi promosi.

Abstract

The dynamics of educational patterns that are so fast and successe make competition between schools more stringent and attracte. , With this competition the right marketing strategy for educational institutions is absolutely necessary, no exception for Muhammadiyah Cimanggu vocational high School. To find out an effecte strategy, this research is done by processing history data with the aim, can provide different information from these data. Data processing methods like this are often referred to as data mining. K-Means is a simple non-hierarchical data clustering method. In this study, comparing two distance spaces is between Manhattan (City Block) and Euclidean. The result of the Research of the writing it can be seen that the best number of clusters in the data of students of Cimanggu Muhammadiyah vocational high School are 3 clusters, this can be seen from the results of the difference in Sum Square Error (SSE), the biggest gap in the Euclidean distance space method is 5.882 located at the point of cluster 3 and the biggest score on the Manhattan distance space method is 5,758 located in cluster 3, this shows that the optimal number of clusters for students of Muhammadiyah Cimanggu is three clusters. seen from Average within Centroid distance, Euclidean is smaller than manhattaan which is $15,115 < 15,398$, therefore the optimal distance measure to be used in the case of clustering data of Muhammadiyah Cimanggu vocational high School is Euclidean distance.

Key Word: k means, *manhattan*, *euclidean*, promotion strategy.



1. Pendahuluan

Dinamika pola pendidikan yang begitu cepat dan silih berganti menjadikan persaingan antar sekolah semakin ketat. Hal ini tentunya menjadi sinyal positif dalam hal peningkatan kualitas penyelenggaraan pendidikan. Dengan adanya persaingan antar sekolah yang semakin atraktif, strategi pemasaran yang tepat untuk lembaga pendidikan mutlak diperlukan, tak terkecuali untuk SMK Muhammadiyah Cimanggu

SMK Muhammadiyah Cimanggu membutuhkan strategi pemasaran yang efektif. Untuk memperoleh strategi pemasaran yang efektif, maka penulisan ini dilakukan dengan cara mengolah data *history* dengan tujuan mengetahui pola, sehingga dapat mengambil informasi yang tersembunyi dari data-data tersebut. Metode pengolahan data seperti ini sering disebut sebagai data mining. Pada penulisan ini analisa data mining dilakukan dengan menggunakan Algoritma K-Means Clustering.

(Ong, 2013) data data yang telah didapatkan dapat dikelompokkan kedalam beberapa *cluster* berdasarkan kemiripan dari data-data tersebut, sehingga data-data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster* dan yang memiliki karakteristik yang berbeda dikelompokkan dalam *cluster* yang lain yang memiliki karakteristik yang sama.

Data yang dibuat dalam cluster yang berbeda tersebut dihitung 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*).

Pada penulisan ini, akan membandingkan dua *distance space* yaitu antara *Manhattan (City Block)* dan *Euclidean*. Dengan adanya perbandingan ini, diharapkan penulis dapat menemukan metode perhitungan jarak yang paling baik, sehingga dapat melakukan pemasaran dengan strategi yang tepat untuk mendapatkan calon siswa baru.

Data Mining

Menurut (Sulastri & Gufroni, 2017) merupakan suatu proses penggalian data atau penyaringan data dengan memanfaatkan kumpulan data dengan ukuran yang cukup besar melalui serangkaian proses untuk

mendapatkan informasi yang berharga dari data tersebut.

(Burhanuddin, Utami, & Pramono, 2017) proses clustering berlangsung jika proses perbandingan satu persatu data selesai hingga kelompok klaster terpenuhi.

Menurut (Anggara, Sujiani, & Helfi, 2016) Clustering merupakan salah satu metode Data Mining yang bersifat tanpa arahan (*unsupervised*). Ada dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu *hierarchical (hirarki)* data clustering dan *non-hierarchical (non hirarki)* data clustering.

K-Means

Menurut (Yudi Agusta, 2007) K-means clustering merupakan salah satu metode data *clustering* non-hirarki yang mengelompokkan data dalam bentuk satu atau lebih *cluster / kelompok*. Data-data yang memiliki karakteristik yang sama dikelompokkan dalam satu *cluster / kelompok* dan data yang memiliki karakteristik yang berbeda dikelompokkan dengan *cluster/kelompok* yang lain sehingga data yang berada dalam satu *cluster/kelompok* memiliki tingkat variasi yang kecil

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

- a. Tentukan jumlah *cluster*
- b. Tentukan k *centroid* (titik pusat *cluster*) awal secara random

Penentuan *centroid* awal dilakukan secara random/acak dari objek objek yang tersedia sebanyak k *cluster*, kemudian untuk menghitung *centroid cluster* ke-i berikutnya, digunakan rumus sebagai berikut :

$$v = \frac{\sum_{i=1}^n x_i}{n} \quad ; i = 1, 2, 3, \dots, n$$

Dimana;

v : *centroid* pada *cluster*

x_i : objek ke-i

n : banyaknya objek/jumlah objek yang menjadi anggota *cluster*

- c. Hitung jarak setiap objek ke masing-masing *centroid* dari masing-masing *cluster*. Alokasikan masing-masing objek ke dalam *centroid* yang paling dekat. Untuk melakukan pengalokasian objek kedalam masing-masing *cluster* pada saat iterasi secara umum dapat dilakukan dengan cara hard k-means dimana secara tegas setiap objek dinyatakan sebagai anggota *cluster* dengan mengukur jarak

- kedekatan sifatnya terhadap titik pusat *cluster* tersebut.
- Lakukan iterasi, kemudian tentukan posisi *centroid* baru dengan menggunakan persamaan (2.2)
 - Ulangi langkah 3 jika posisi *centroid* baru tidak sama

Distance Space Untuk Menghitung Jarak Antara Data dan Centroid

Beberapa *distance space* telah 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 dua titik x_1 dan x_2 pada *Manhattan/City Block distance space* dihitung dengan menggunakan rumus sebagai berikut (Yudi Augusta, 2007)

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

Sedangkan untuk L2 (*Euclidean distance space*), jarak antara dua titik dihitung menggunakan rumus sebagai berikut:

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

Dimana :

D_{L_2} = jarak kuadrat *Euclidean* antar objek ke x_2 dengan x_1 .

P = jumlah variabel *cluster*.

x_{2j} = nilai atau data dari objek ke-2 pada variabel ke- j .

x_{1j} = nilai atau data dari objek ke-1 pada variabel ke- j (Everitt, 1993).

Metode Elbow

Menurut Bholowalia et al dalam (Eka Merliana, Ernawati, & Santoso, 2015) menjelaskan bahwa "Metode *Elbow* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik".

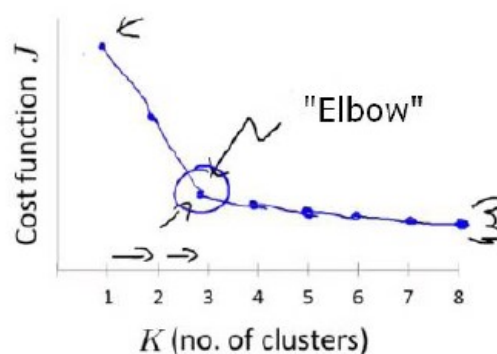
Metode ini memberikan ide/gagasan dengan cara memilih nilai *cluster* dan kemudian menambah nilai *cluster* tersebut untuk dijadikan model data dalam penentuan *cluster* terbaik. Dan selain itu persentase perhitungan yang dihasilkan menjadi perbandingan antara jumlah *cluster* yang ditambah. Hasil persentase yang berbeda dari

setiap nilai *cluster* dapat ditunjukkan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka nilai *cluster* tersebut yang terbaik .

Untuk mendapatkan perbandingannya adalah dengan menghitung SSE (*Sum of Square Error*) dari masing-masing nilai *cluster*. Karena semakin besar jumlah *cluster* K maka nilai SSE akan semakin kecil. Rumus SSE pada K-Means

$$SSE = \sum_{K=1}^K \sum_{x_j \in S_K} \|x_j - c_k\|_2^2$$

Setelah dilihat akan ada beberapa nilai K yang mengalami penurunan paling besar dan selanjutnya hasil dari nilai K akan turun secara perlahan-lahan sampai hasil dari nilai K tersebut stabil. Misalnya nilai *cluster* K=2 ke K=3, kemudian dari K=3 ke K=4, terlihat penurunan drastis membentuk siku pada titik K=3 maka nilai *cluster* k yang ideal adalah K=3.



Sumber: (Eka Merliana et al., 2015)

Gambar 1. Grafik Elbow

Algoritma Metode *Elbow* dalam menentukan nilai K pada K-Means

- Mulai
 - Inisialisasi awal nilai K
 - Naikkan nilai K
 - Hitung hasil *sum of square error* dari tiap nilai K
 - Melihat hasil *sum of square error* dari nilai K yang turun secara drastis
 - Tetapkan nilai K yang beberbentuk siku
- Selesai

Promosi

Menurut (Hedynata & E.D.Radianto, 2016) Strategi promosi berjalan kurang maksimal disebabkan oleh ketidaktahuan LUSCIOUS mengenai strategi promosi mana

yang sebenarnya paling dapat meningkatkan penjualan dan jumlah konsumen, sehingga mengakibatkan perusahaan tidak fokus dan akhirnya mengandalkan semua cara dan media promosi yang dimiliki

Penulisan Terkait

- a. Penulisan yang dilakukan oleh (Hendrawan, 2015) berjudul "Merancang Sistem Strategi Promosi Sekolah Dengan Metode K-Means *Clustering* Di SMKN 1 Ngasem Kediri". Penulisan ini dilatarbelakangi karena promosi sekolah SMK Negeri 1 Ngasem masih menggunakan sistem manual dengan mengirimkan panitia PPDB untuk menyerahkan surat penerimaan calon peserta didik baru ke SMP yang sudah ditentukan, penyebaran pamflet di tempat secara acak tidak terorganisir sehingga tidak efektif dan efisien, maka dibangun aplikasi sistem promosi sekolah menggunakan *K-Means Clustering* guna untuk mempermudah admin dalam menyusun strategi promosi sekolah. Hasil dari penulisan ini adalah pada *cluster* pertama didominasi oleh siswa dari jurusan teknik komputer dan jaringan yang mempunyai nilai akademik rata-rata 170,887 dan berada di daerah sekitar kecamatan Plemahan, Tarokan, Badas. *Cluster* kedua didominasi jurusan Teknik Komputer jaringan dan jasa boga dengan nilai akademik rata-rata 169,697 yang berasal dari kecamatan Gampengrejo, Pagu, Banyak, Kandangan. *Cluster* ketiga didominasi oleh siswa jurusan Teknik Kendaraan Ringan dan Teknik Gambar bangunan dengan nilai akademik rata-rata 158,509 yang berasal dari kecamatan sekitar Gurah, Kayen kidul, Plosoklaten, Pesantren. Kata Kunci : Promosi, K-Means *Clustering*, SMK Negeri 1 Ngasem Kediri.
- b. Penulisan yang dilakukan oleh (Chasanah & Widiyono, 2017) berjudul "Penentuan Strategi Promosi Penerimaan Mahasiswa Baru Dengan Algoritma *Clustering K-Means*". Setiap pelaksanaan, terdapat beberapa kendala dan permasalahan. sejak tiga tahun yang lalu terjadi penurunan jumlah pendaftar calon mahasiswa baru. Beberapa evaluasi dilakukan untuk meningkatkan jumlah pendaftar mahasiswa baru. Permasalahan lainnya yaitu kurangnya peminat program studi tertentu, sehingga pendaftar di program studi Akuntansi Komputer sedikit. Berdasarkan data pendaftaran siswa baru dalam penulisan ini akan dianalisa dengan metode pengelompokan / klustering menggunakan algoritma K-Means untuk mendapatkan informasi-informasi sesuai kelompok-kelompok tertentu dari data pendaftaran tahun sebelumnya. Hasil klustering akan digunakan untuk menentukan strategi promosi penerimaan mahasiswa baru STMIK Widya Pratama Pekalongan. Kata
- c. Penulisan oleh (Yoga Religia, 2016) dengan judul "Metode Manhattan, Euclidean Dan Chebyshev Pada Algoritma K-Means Untuk Pengelompokan Status Desa". Kementerian Desa, Pembangunan Daerah Tertinggal dan Transmigrasi (Kementerian Desa PDTT) merupakan kementerian dalam Pemerintahan Indonesia yang membidangi urusan pembangunan desa dan kawasan perdesaan, pemberdayaan masyarakat desa, percepatan pembangunan daerah tertinggal, dan transmigrasi. Data Potensi Desa tahun 2014 (Podes 2014) merupakan data yang dikeluarkan oleh Badan Pusat Statistik yang bekerjasama dengan Kementerian Desa PDTT dalam bentuk unsupervised dan terdiri dari 74093. Data Podes 2014 dibuat berdasarkan tingkat perkembangan desa (*village specific*) di Indonesia dengan menjadikan desa sebagai unit analisis. Dalam penulisan ini akan dilakukan perbandingan metode perhitungan jarak pada k-means antara Manhattan, Euclidean dan Chebychev. Pengujian akan dilakukan menggunakan waktu eksekusi dan *davies bouldin index*. Dari pengujian yang sudah dilakukan, metode perhitungan jarak Chebyshev memiliki akumulasi waktu paling efisien dibandingkan Manhattan dan Euclidean, sedangkan metode Euclidean memiliki nilai *Index Davies* paling optimal dibanding metode Manhattan dan Chebyshev. Kata kunci: Pembangunan Desa, k-means, *Manhattan*, *Euclidean*, *Chebychev*

2. Metode Penulisan

Penentuan jumlah kluster menggunakan metode elbow lebih baik jika dibandingkan dengan menentukan jumlah cluster sesuai kehendak. Penulis berharap dengan menggunakan metode elbow ini Jumlah cluster yang dibentuk merupakan jumlah cluster terbaik dari data yang ada.

Perbandingan Hasil clustering data menggunakan metode K-Means dengan

rumus distance space antara *Manhattan(City Block)* dan *Euclidean* bertujuan untuk mengetahui perbedaan pola antara kedua *distance space* tersebut dalam melakukan pengelompokan data, sehingga dapat dipelajari pola pengelompokan data untuk masing-masing rumus *distance space* dengan tujuan mencari pola terbaik diantara dua rumus distance space tersebut.

Tahapan Penulisan

- Pengumpulan dataset Tahapan pertama dalam penulisan ini adalah melakukan pengumpulan data yang merupakan hasil dari riset yang dilakukan penulis. Data set yang diperoleh adalah data diri siswa.
- Pengolahan dataset Data pada penulisan ini sebelum digunakan untuk eksperimen dan penulisan terlebih dahulu perlu dilakukan pengolahan data adalah Normalisasi data dan Pemberian bobot nilai agar dapat dihitung secara numerik
- Melakukan analisis hasil eksperimen.
- Melakukan evaluasi hasil eksperimen.

Metode Pengumpulan data

Jenis penulisan yang digunakan dalam penulisan ini adalah jenis penulisan kuantitatif, yaitu melakukan analisis berdasarkan pada data-data numerik (angka) dari data siswa yang ada di SMK Muhammadiyah Cimanggu dengan data yang dikumpulkan melalui wawancara, observasi dan studi kepustakaan.

Populasi dan Sample

Populasi yang penulis gunakan dalam penulisan ini adalah 351 data siswa SMK Muhammadiyah cimanggu.

Metode Analisis Data

Untuk menganalisis data dalam penerapan data mining ini menggunakan proses tahapan *knowledge discovery in databases* (KDD) yang terdiri dari *Database*, *Data Cleaning*, *Data integration*, *Task relevan data*, *Data transformation*, *Data mining*, *Pattern evolution*, *knowledge*:

3. Hasil Dan Pembahasan

3.1. Perhitungan Manual Metode K-Means

Penerapan Metode klusterisasi K-Means menggunakan perhitungan manual dilakukan penulis pada 351 data siswa SMK Muhammadiyah Cimanggu. Sesuai dengan judul yang diambil penulis melakukan

perhitungan dengan menggunakan *Euclidean distance space* dan *manhattan distance space*, dimana dalam proses perhitungan manual ini penulis mengambil jumlah *cluster* sebanyak 3 *cluster*.

Berikut ini merupakan langkah perhitungan metode K-Means menggunakan *Euclidean Distance space*:

- Jumlah *cluster* yang ditentukan adalah 3 *cluster*
- k *Centroid* (titik pusat *cluster*) awal secara random yang dipilih penulis adalah sebagai berikut :
Centroid 1 (M1)=(3,3,3)
Centroid 2 (M2)=(3,6,5)
Centroid 3 (M3)=(2,5,1).
- Perhitungan jarak setiap objek ke masing-masing *Centroid* dari masing-masing *cluster* menggunakan *Euclidean distance space* untuk iterasi pertama. Menggunakan rumus:

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

$$DM1 = \sqrt{(2-3)^2 + (4-3)^2 + (3-3)^2} = 1,41421356$$

$$DM2 = \sqrt{(2-3)^2 + (4-6)^2 + (3-5)^2} = 3$$

$$DM3 = \sqrt{(2-2)^2 + (4-5)^2 + (3-1)^2} = 2,23606798$$

Berikut ini tabel hasil perhitungan jarak antara masing-masing *cluster*, pada tabel dibawah ini penulis hanya menampilkan 10 data pertama.

Tabel 1. Jarak *Cluster* menggunakan *Euclidean* iterasi 1

| N M | J R | A L | S H | Jarak Ke Centroid cluster 1 | Jarak Ke Centroid cluster 2 | Jarak Ke Centroid cluster 3 | Jarak Terdekat | C |
|-----|-----|-----|-----|-----------------------------|-----------------------------|-----------------------------|----------------|-----|
| A | 2 | 4 | 3 | 1,41421356 | 3 | 2,23606798 | 1,41421356 | c 1 |
| B | 2 | 3 | 2 | 1,41421356 | 4,35889894 | 2,23606798 | 1,41421356 | c 1 |
| C | 2 | 6 | 17 | 14,3527001 | 12,0415946 | 16,0312195 | 12,04159458 | c 2 |
| D | 2 | 21 | 30 | 32,4653662 | 29,1719043 | 33,1209903 | 29,17190429 | c 2 |
| E | 2 | 1 | 5 | 3 | 5,09901951 | 5,65685425 | 3 | c 1 |
| F | 2 | 5 | 5 | 3 | 1,41421356 | 4 | 1,41421356 | c 2 |
| G | 2 | 2 | 2 | 1,73205081 | 5,09901951 | 3,16227766 | 1,732050808 | c 1 |
| H | 2 | 5 | 5 | 3 | 1,41421356 | 4 | 1,41421356 | c 2 |
| I | 2 | 6 | 5 | 3,74165739 | 1 | 4,12310563 | 1 | c 2 |
| J | 2 | 3 | 3 | 1 | 3,74165739 | 2,82842712 | 1 | c 1 |

Dari Hasil Perhitungan secara keseluruhan didapatkan keanggotaan seperti tampak pada table 2.

Tabel 2. Hasil *Clustering* menggunakan *Euclidean* iterasi 1

| Cluster | Jumlah Anggota |
|---------------|----------------|
| C0 | 191 |
| C1 | 106 |
| C2 | 54 |
| Total Anggota | 351 |

Setelah selesai melakukan pengelompokan harus dihitung pula rasio antara BCV(Between Cluster Variation) dengan WCV(Within Cluster Variation):

$$d(m_1, m_2) = \sqrt{(3-3)^2 + (3-6)^2 + (3-5)^2} = 3,60555128$$

$$d(m_1, m_3) = \sqrt{(3-2)^2 + (3-5)^2 + (3-1)^2} = 3$$

$$d(m_2, m_3) = \sqrt{(3-2)^2 + (6-5)^2 + (5-1)^2} = 4,24264069$$

$$\begin{aligned} BCV &= d(m_1, m_2) + d(m_1, m_3) + d(m_2, m_3) \\ &= 3,60555128 + 3 + 4,24264069 \\ &= 10,848192 \end{aligned}$$

Perhitungan WCV diambil dari hasil jumlah pangkat dua dari jarak terdekat

Tabel 3. Jarak Terdekat Menggunakan *Euclidean* iterasi 1

| Jarak Terdekat | Pangkat 2 dari jarak terdekat |
|----------------|-------------------------------|
| 1,414213562 | 2 |
| 1,414213562 | 2 |
| 12,04159458 | 145 |
| 29,17190429 | 851 |
| 3 | 9 |
| 1,414213562 | 2 |
| 1,732050808 | 3 |
| 1,414213562 | 2 |
| 1 | 1 |
| 1 | 1 |
| N | |

$$WCV = 2+2+145+851+9+2+3+2+1+1+...=14344$$

$$\begin{aligned} \text{Besaran Rasio} &= BCV/WCV \\ &= 10,848192/14344 \\ &= 0,00076 \end{aligned}$$

- d. Lakukan iterasi, kemudian tentukan posisi *Centroid* baru
Pembaruan nilai *Centroid* dengan menghitung rata-rata nilai pada masing-masing *cluster*.

Tabel 4.

Jumlah Nilai Pada *Euclidean* iterasi 1

| Cluster | Total jurusan | Total alamat | Total sekolah | Jumlah siswa |
|-------------|---------------|--------------|---------------|--------------|
| c1 | 362 | 397 | 485 | 191 |
| c2 | 189 | 628 | 1272 | 106 |
| c3 | 87 | 286 | 69 | 54 |
| Grand Total | 638 | 1311 | 1826 | 351 |

Tabel 5.

Rata-rata Nilai Pada *Euclidean* iterasi

| Mean Of Jurusan | Mean Off Alamat | Mean Off Sekolah |
|-----------------|-----------------|------------------|
| 1,895287958 | 2,078534031 | 2,539267016 |
| 1,783018868 | 5,924528302 | 12 |
| 1,611111111 | 5,296296296 | 1,277777778 |

Dari hasil perhitungan diperoleh *Centroid* baru yaitu :

$$M1 = (1,895287958, 2,078534031, 2,539267016)$$

$$M2 = (1,783018868, 5,924528302, 12)$$

$$M3 = (1,611111111, 5,296296296, 1,277777778)$$

Setelah mendapatkan *Centroid* baru langkah selanjutnya mengulang kembali perhitungan pada langkah 4 yaitu perhitungan jarak setiap objek ke masing-masing *Centroid* dari masing-masing *cluster* menggunakan *Euclidean distance space* untuk iterasi kedua.

$$\begin{aligned} DM1 &= \sqrt{(2-1,895287958)^2 + (4-2,078534031)^2 + (3-2,539267016)^2} \\ &= 1,978704365 \end{aligned}$$

$$\begin{aligned} DM2 &= \sqrt{(2-1,783018868)^2 + (4-5,924528302)^2 + (3-12)^2} \\ &= 9,206024658 \end{aligned}$$

$$\begin{aligned} DM3 &= \sqrt{(2-1,611111111)^2 + (4-5,296296296)^2 + (3-1,277777778)^2} \\ &= 2,19035797 \end{aligned}$$

Berikut ini tabel hasil perhitungan jarak untuk iterasi kedua, dimana pada tabel dibawah ini penulis hanya menampilkan 10 data pertama

Tabel 6. Jarak *Cluster* menggunakan *Euclidean* iterasi 1

| N | J | A | S | Jarak Ke Centroid cluster 1 | Jarak Ke Centroid cluster 2 | Jarak Ke Centroid cluster 3 | Jarak Terdekat | C |
|---|---|---|---|-----------------------------|-----------------------------|-----------------------------|----------------|----|
| A | 2 | 4 | 3 | 1,978704365 | 9,206024658 | 2,19035797 | 1,978704365 | c1 |

| | | | | | | | | |
|---|---|--------|--------|-----------------|-----------------|-----------------|-------------------------|--------|
| B | 2 | 3 | 2 | 1,0727 87517 | 10,421 12981 | 2,43840 4435 | 1,07 278 751 7 | c 1 |
| C | 2 | 6 | 1 7 | 14,983 37941 | 5,0052 74896 | 15,7427 6675 | 5,00 527 489 6 | c 1 |
| D | 2 | 2 1 | 3 0 | 33,348 53363 | 23,480 13901 | 32,7371 8977 | 23,4 801 390 1 | c 3 |
| E | 2 | 1 | 5 | 2,6887 55677 | 8,5614 28608 | 5,69774 8214 | 2,68 875 567 7 | c 1 |
| F | 2 | 5 | 5 | 3,8211 43132 | 7,0641 22974 | 3,75419 2901 | 3,75 419 290 1 | c 1 |
| G | 2 | 2 | 2 | 0,5549 24427 | 10,744 71978 | 3,39682 3336 | 0,55 492 442 7 | c 2 |
| H | 2 | 5 | 5 | 3,8211 43132 | 7,0641 22974 | 3,75419 2901 | 3,75 419 290 1 | c 2 |
| I | 2 | 6 | 5 | 4,6307 73885 | 7,0037 68756 | 3,80806 6667 | 3,80 806 666 7 | c 1 |
| J | 2 | 3 | 3 | 1,0355 38037 | 9,4657 24832 | 2,89659 4661 | 1,03 553 803 7 | c 3 |

Dari Hasil Perhitungan secara keseluruhan didapatkan keanggotaan seperti terlihat pada table 7.

Tabel 7. Hasil *Clustering* menggunakan *Euclidean* iterasi 2

| Cluster | Jumlah Anggota |
|-------------|----------------|
| C0 | 199 |
| C1 | 70 |
| C2 | 82 |
| Grand Total | 351 |

Setelah selesai melakukan pengelompokan harus dihitung pula rasio antara *BCV* (*Between Cluster Variation*) dengan *WCV* (*Within Cluster Variation*) untuk iterasi ke-2 :

$$d(m1,m2)=\sqrt{(1,895287958-1,783018868)^2+(2,078534031-5,924528302)^2+(2,539267016-12)^2}$$

$$=10,21321423$$

$$d(m1,m3)=\sqrt{(1,895287958-1,611111111)^2+(2,078534031-5,296296296)^2+(2,539267016-1,277777778)^2}$$

$$=3,467867583$$

$$d(m2,m3)=\sqrt{(1,783018868-1,611111111)^2+(5,924528302-5,296296296)^2+(12-1,277777778)^2}$$

$$=10,74198665$$

$$BCV=d(m1,m2)+d(3,467867583+10,74198665=24,42306846$$

Perhitungan *WCV* diambil dari hasil jumlah pangkat dua dari jarak terdekat pada iterasi kedua.

Tabel 8. Jarak Terdekat Menggunakan *Euclidean* iterasi 2

| Jarak Terdekat | Pangkat 2 dari jarak terdekat |
|----------------|-------------------------------|
| 1,978704365 | 3,915270963 |
| 1,072787517 | 1,150873057 |
| 5,005274896 | 25,05277679 |
| 23,48013901 | 551,3169277 |
| 2,688755677 | 7,229407089 |
| 3,754192901 | 14,09396433 |
| 0,554924427 | 0,30794112 |
| 3,754192901 | 14,09396433 |
| 3,808066667 | 14,50137174 |
| 1,035538037 | 1,072339026 |
| N | ... |

$$WCV=3,915270963+1,150873057+25,05277679+551,3169277+7,229407089+14,09396433+0,30794112+14,09396433+14,50137174+1,072339026+...=7626,292021$$

$$\text{Besaran Rasio} = BCV/WCV = 24,42306846/7626,292021 = 0,00206$$

e. Ulangi langkah 3 jika posisi *Centroid* baru tidak sama

Berdasarkan hasil perhitungan nilai rasio pada iterasi pertama dan iterasi berbeda maka iterasi dilanjutkan ke iterasi berikutnya, penulis menemukan nilai rasio yang sama dengan keanggotaan yang tidak berubah ditemukan pada iterasi ke 19 dimana nilai rasio iterasi 18 dan iterasi 19 bernilai sama yaitu 0,0004, sehingga proses perulangan dihentikan pada iterasi ke 19. Berikut ini merupakan hasil akhir pengklasteran yang dibentuk dengan jumlah nilai k=3

Tabel 9. Hasil Akhir *Clustering* menggunakan *Euclidean*

| Cluster | Jumlah Anggota |
|---------|----------------|
| C0 | 109 |
| C1 | 182 |
| C2 | 60 |
| Total | 351 |

Berikut ini merupakan langkah perhitungan metode K-Means menggunakan *Manhattan distance space*:

- Jumlah *cluster* yang ditentukan adalah 3 *cluster*
- k *Centroid* (titik pusat *cluster*) awal secara random yang dipilih penulis adalah sebagai berikut :
Centroid 1 (M1)=(3,3,3)
Centroid 2 (M2)=(3,6,5)
Centroid 3 (M3)=(2,5,1)
- Perhitungan jarak setiap objek ke masing-masing *Centroid* dari masing-masing *cluster* menggunakan *Euclidean distance space* untuk iterasi pertama. Menggunakan rumus

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

$$DM1 = |2-3| + |4-3| + |3-3| = 2$$

$$DM2 = |2-3| + |4-6| + |3-5| = 5$$

$$DM3 = |2-2| + |4-5| + |3-1| = 3$$

Berikut ini tabel hasil perhitungan jarak antara masing-masing, pada tabel dibawah ini penulis hanya menampilkan 10 data pertama untuk data lengkap dapat dilihat di lampiran.

Tabel 10. Jarak *Cluster* menggunakan *Manhattan* iterasi 1

| N M | J R | A L | S H | Jarak Ke Centroid cluster 1 | Jarak Ke Centroid cluster 2 | Jarak Ke Centroid cluster 3 | Jarak Terdekat | Cluster |
|-----|-----|-----|-----|-----------------------------|-----------------------------|-----------------------------|----------------|---------|
| A | 2 | 4 | 3 | 2 | 5 | 3 | 2 | c1 |
| B | 2 | 3 | 2 | 2 | 7 | 3 | 2 | c1 |
| C | 2 | 6 | 1 | 18 | 13 | 17 | 13 | c2 |
| D | 2 | 2 | 3 | 46 | 41 | 45 | 41 | c2 |
| E | 2 | 1 | 5 | 5 | 6 | 8 | 5 | c1 |
| F | 2 | 5 | 5 | 5 | 2 | 4 | 2 | c2 |
| G | 2 | 2 | 2 | 3 | 8 | 4 | 3 | c1 |
| H | 2 | 5 | 5 | 5 | 2 | 4 | 2 | c2 |
| I | 2 | 6 | 5 | 6 | 1 | 5 | 1 | c2 |
| J | 2 | 3 | 3 | 1 | 6 | 4 | 1 | c1 |

Dari Hasil Perhitungan secara keseluruhan didapatkan keanggotaan seperti terlihat pada table 11.

Tabel 11. Hasil *Clustering* menggunakan *Manhattan* iterasi 1

| Cluster | Jumlah Anggota |
|---------------|----------------|
| C0 | 196 |
| C1 | 64 |
| C2 | 91 |
| Total Anggota | 351 |

Setelah selesai melakukan pengelompokan dihitung pula rasio antara BCV(Between *Cluster Variation*) dengan WCV(Within *Cluster Variation*):

$$d(m1, m2) = |3-3| + |3-6| + |3-5| = 5$$

$$d(m1, m3) = |3-2| + |3-5| + |3-1| = 5$$

$$d(m2, m3) = |3-2| + |6-5| + |5-1| = 6$$

$$BCV = d(m1, m2) + d(m1, m3) + d(m2, m3) = 5 + 5 + 6 = 16$$

Perhitungan WCV diambil dari hasil jumlah pangkat dua dari jarak terdekat

Tabel 12. Jarak Terdekat Menggunakan *Manhattan* iterasi 1

| Jarak Terdekat | $\wedge 2$ dari jarak min |
|----------------|---------------------------|
| 2 | 4 |
| 2 | 4 |
| 13 | 169 |
| 41 | 1681 |
| 5 | 25 |
| 2 | 4 |
| 3 | 9 |
| 2 | 4 |
| 1 | 1 |
| 1 | 1 |
| Jarak ke-n | ... |

$$WCV = 4+4+169+1681+25+4+9+4+1+1+...=23777$$

$$\text{Besaran Rasio} = BCV/WCV$$

$$= 16/23777$$

$$= 0,00067$$

- Lakukan iterasi, kemudian tentukan posisi *Centroid* baru

Pembaruan nilai *Centroid* dengan menghitung rata-rata nilai pada masing-masing *cluster*.

Tabel 13. Jumlah Nilai Pada *Manhattan* iterasi 1

| Row Labels | Sum Of Jurusan | Sum Of Alamat | Sum Of Sekolah | Count Of Nama |
|-------------|----------------|---------------|----------------|---------------|
| C0 | 390 | 429 | 1007 | 196 |
| C1 | 107 | 523 | 713 | 64 |
| C2 | 141 | 359 | 106 | 91 |
| Grand Total | 638 | 1311 | 1826 | 351 |

Tabel 14. Rata-rata Nilai Pada *Manhattan* iterasi 1

| Mean Of Jurusan | Mean Off Alamat | Mean Off Sekolah |
|-----------------|-----------------|------------------|
| 1,989795918 | 2,18877551 | 5,137755102 |
| 1,671875 | 8,171875 | 11,140625 |
| 1,549450549 | 3,945054945 | 1,164835165 |

Dari hasil perhitungan diperoleh *Centroid* baru yaitu :

$$M1=(1,989795918, 2,18877551, 5,137755102)$$

$$M2=(1,671875, 8,171875, 11,140625)$$

$$M3=(1,549450549, 3,945054945, 1,164835165)$$

Setelah mendapatkan *Centroid* baru langkah selanjutnya mengulang kembali perhitungan pada langkah 4 yaitu perhitungan jarak setiap objek ke masing-masing *Centroid* dari masing-masing *cluster* menggunakan *manhattan distance space* untuk iterasi kedua.

$$DM1=|2-1,989795918| + |4-2,18877551| + |3-5,137755102| = 3,959183673$$

$$DM2=|2-1,671875| + |4-8,171875| + |3-11,140625| = 12,640625$$

$$DM3=|2-1,549450549| + |4-3,945054945| + |3-11,164835165| = 2,340659341$$

Berikut ini tabel hasil perhitungan jarak untuk iterasi kedua, dimana pada tabel dibawah ini penulis hanya menampilkan 10 data pertama.

Tabel 15. Jarak *Cluster* menggunakan *Manhattan* iterasi 2

| N M | J R | A L | S H | Jarak Ke Centroid cluster 1 | Jarak Ke Centroid cluster 2 | Jarak Ke Centroid cluster 3 | Jarak Terdekat | C |
|-----|-----|-----|-----|-----------------------------|-----------------------------|-----------------------------|----------------|----|
| A | 2 | 4 | 3 | 3,959183673 | 12,640625 | 2,340659341 | 2,340659341 | c1 |
| B | 2 | 3 | 2 | 3,959183673 | 14,640625 | 2,230769231 | 2,230769231 | c1 |
| C | 2 | 6 | 1 | 15,68367347 | 8,359375 | 18,340659341 | 8,359375 | c1 |
| D | 2 | 2 | 3 | 43,68367347 | 32,015625 | 46,340659341 | 32,015625 | c3 |
| E | 2 | 1 | 5 | 1,336734694 | 13,640625 | 7,230769231 | 1,336734694 | c3 |
| F | 2 | 5 | 5 | 2,959183673 | 9,640625 | 5,340659341 | 2,959183673 | c3 |
| G | 2 | 2 | 2 | 3,336734694 | 15,640625 | 3,230769231 | 3,230769231 | c2 |
| H | 2 | 5 | 5 | 2,959183673 | 9,640625 | 5,340659341 | 2,959183673 | c2 |
| I | 2 | 6 | 5 | 3,959183673 | 8,640625 | 6,340659341 | 3,959183673 | c1 |
| J | 2 | 3 | 3 | 2,959183673 | 13,640625 | 3,230769231 | 2,959183673 | c1 |

Dari Hasil Perhitungan secara keseluruhan didapatkan keanggotaan sebagai berikut:

Tabel 16. Hasil *Clustering* menggunakan *Manhattan* iterasi 2

| Cluster | Jumlah Anggota |
|-------------|----------------|
| C0 | 116 |
| C1 | 46 |
| C2 | 189 |
| Grand Total | 351 |

Setelah selesai melakukan pengelompokan harus dihitung pula rasio antara *BCV*(Between Cluster Variation) dengan *WCV*(Within Cluster Variation) untuk iterasi ke-2 :

$$d(m1,m2)= |1,989795918-1,671875| + |2,18877551-8,171875| + |5,137755102-11,140625| = 12,30389031$$

$$d(m1,m3)= |1,989795918-1,549450549| + |2,18877551-3,945054945| + |5,137755102-1,164835165| = 6,169544741$$

$$d(m2,m3)= |1,671875-1,549450549| + |8,171875-3,945054945|^2 + |11,140625-1,164835165| = 14,32503434$$

$$BCV= d(m1,m2)+ d(m1,m3)+ d(m2,m3)$$

$$= 12,30389031+6,169544741+14,32503434 = 32,79846939$$

Perhitungan WCV diambil dari hasil jumlah pangkat dua dari jarak terdekat pada ierasi kedua.

Tabel 17. Jarak Terdekat Menggunakan *Manhattan* iterasi 2

| Jarak Terdekat | \wedge^2 dari jarak min |
|----------------|---------------------------|
| 2,340659341 | 5,478686149 |
| 2,230769231 | 4,976331361 |
| 8,359375 | 69,87915039 |
| 32,015625 | 1025,000244 |
| 1,336734694 | 1,786859642 |
| 2,959183673 | 8,756768013 |
| 3,230769231 | 10,43786982 |
| 2,959183673 | 8,756768013 |
| 3,959183673 | 15,67513536 |
| 2,959183673 | 8,756768013 |
| Data ke n | |

WCV=5,478686149+4,976331361+69,87915039+1025,000244+1,786859642+8,756768013+10,43786982+8,756768013+15,67513536+8,756768013+...=14948,95541

Besaran Rasio= BCV/WCV
=32,79846939/14948,95541
=0,00219

Ulangi langkah 3 jika posisi *Centroid* baru tidak sama

Berdasarkan hasil perhitungan nilai rasio pada iterasi pertama dan iterasi kedua berbeda maka iterasi dilanjutkan ke iterasi berikutnya, penulis menemukan nilai rasio yang sama dengan keanggotaan yang tidak berubah ditemukan pada iterasi ke 8 dimana nilai rasio iterasi 7 dan iterasi 8 bernilai sama yaitu 0,0003, sehingga proses perulangan dihentikan pada iterasi ke 8. Berikut ini merupakan hasil akhir pengklusteran yang dibentuk dengan jumlah nilai k=3

Tabel 18. Hasil Akhir *Clustering* menggunakan *Manhattan*

| Row Labels | Count of nama |
|-------------|---------------|
| C0 | 184 |
| C1 | 118 |
| C2 | 49 |
| Grand Total | 351 |

Implementasi K-Means Menggunakan Rapidminer

Penulis melakukan eksperimen menggunakan aplikasi rapidminer dalam proses klusterisasi menggunakan metode k-means, dimana eksperimen dilakukan sebanyak 5 kali pada masing-masing distance. Implementasi pertama dilakukan dengan menggunakan *Euclidean distance space*, implementasi kedua dilakukan dengan menggunakan *Manhattan distance space* masing-masing dilakukan sebanyak 5 kali percobaan yaitu dengan mencoba nilai k=2 sampai dengan nilai k=6.



Gambar 1. Implementasi K-Means Pada Rapidminer

Hasil eksperimen menggunakan *Euclidean distance space* dapat dilihat pada tabel dibawah ini:

Tabel 19. Hasil *Clustering* k=2 (*Euclidean distance space*) Pada Rapidminer

| Atribut | Cluster 0 | Cluster 1 |
|-------------|-----------|-----------|
| Jurusan | 1.685 | 1.842 |
| Alamat | 5.741 | 3.370 |
| Sekolah | 17.185 | 3.024 |
| Jml Anggota | 54 | 297 |

Tabel 20. Hasil *Clustering* k=3 (*Euclidean distance space*) Pada Rapidminer

| Atribut | Cluster 0 | Cluster 1 | Cluster 2 |
|-------------|-----------|-----------|-----------|
| Jurusan | 1.730 | 1.835 | 1.909 |
| Alamat | 4.878 | 3.157 | 6.591 |
| Sekolah | 9.865 | 2.275 | 23.455 |
| Jml anggota | 74 | 225 | 22 |

Tabel 21. Hasil *Clustering* k=4 (*Euclidean distance space*) Pada Rapidminer

| Atribut | Clust er 0 | Clust er 1 | Cluster 2 | Cluster 3 |
|-------------|------------|------------|-----------|-----------|
| Jurusan | 1.614 | 1.944 | 1.835 | 2 |
| Alamat | 3.877 | 9.056 | 2.665 | 7.438 |
| Sekolah | 12.018 | 4.500 | 2.355 | 25.562 |
| Jml anggota | 57 | 36 | 242 | 16 |

Tabel 22. Hasil *Clustering* k=5 (*Euclidean distance space*) Pada Rapidminer

| Atribut | Clus ter 0 | Cluster 1 | Clus ter 2 | Clus ter 3 | Clus ter 4 |
|-------------|------------|--------------------|------------|------------|------------|
| Jurusan | 1.714 | 1.8333333333333333 | 2.077 | 1.814 | 1.738 |
| Alamat | 7.179 | 2.2424242424242424 | 6.231 | 7.086 | 2.119 |
| Sekolah | 15.071 | 2.3181818181818181 | 27 | 2.900 | 9.310 |
| Jml anggota | 28 | 198 | 13 | 70 | 42 |

Tabel 23. Hasil *Clustering* k=6 (*Euclidean distance space*) Pada Rapidminer

| Atribut | Clus ter 0 | Clus ter 1 | Clus ter 2 | Clus ter 3 | Clus ter 4 | Clus ter 5 |
|-------------|------------|------------|------------|------------|------------|------------|
| Jurusan | 1.769 | 1.86875 | 1.783 | 2.077 | 1.747 | 1.738 |
| Alamat | 6.846 | 1.825 | 10.261 | 6.231 | 5 | 2.119 |
| Sekolah | 15.462 | 2.269 | 5.739 | 27 | 2.149 | 9.310 |
| Jml anggota | 26 | 160 | 23 | 13 | 87 | 42 |

Implementasi kedua dilakukan dengan menggunakan *Manhattan distance space* sebanyak 5 kali percobaan yaitu dengan mencoba nilai k=2 sampai dengan nilai k=6 hasil yang diperoleh adalah sebagai berikut:

Tabel 24. Hasil *Clustering* k=2 (*Manhattan*) Pada Rapidminer

| Atribut | Cluster 0 | Cluster 1 |
|-------------|-----------|-----------|
| Jurusan | 1.825 | 1.771 |
| Alamat | 3.356 | 6.125 |
| Sekolah | 3.182 | 17.958 |
| Jml anggota | 303 | 48 |

Tabel 25. Hasil *Clustering* k=3 (*Manhattan*) Pada Rapidminer

| Atribut | Cluster 0 | Cluster 1 | Cluster 2 |
|-------------|-----------|-----------|-----------|
| Jurusan | 1.753 | 2 | 1.816 |
| Alamat | 4.896 | 6.300 | 3.053 |
| Sekolah | 8.597 | 21.333 | 2.148 |
| Jml anggota | 77 | 30 | 244 |

Tabel 26. Hasil *Clustering* k=4 (*Manhattan*) Pada Rapidminer

| Atribut | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 |
|-------------|-----------|-----------|-----------|-----------|
| Jurusan | 1.412 | 1.898 | 2.067 | 1.814 |
| Alamat | 4.706 | 2.214 | 6.600 | 7.086 |
| Sekolah | 12.7253 | 2.712 | 26.067 | 2.900 |
| Jml anggota | 51 | 215 | 15 | 70 |

Tabel 27. Hasil *Clustering* k=5 (*Manhattan*) Pada Rapidminer

| Atribut | Clus ter 0 | Clus ter 1 | Clus ter 2 | Cluster 3 | Cluster 4 |
|-------------|------------|------------|------------|-----------|-----------|
| Jurusan | 1.538 | 1.883 | 2.077 | 1.769 | 1.783 |
| Alamat | 9.731 | 1.833 | 6.231 | 5.606 | 2.109 |
| Sekolah | 13.077 | 2.315 | 27 | 2.625 | 10.587 |
| Jml anggota | 26 | 162 | 13 | 104 | 46 |

Tabel 28. Hasil *Clustering* k=6 (*Manhattan*) Pada Rapidminer

| Atribut | Clus ter 0 | Clus ter 1 | Clus ter 2 | Clus ter 3 | Clus ter 4 | Clus ter 5 |
|-------------|------------|------------|------------|------------|------------|------------|
| Jurusan | 2.077 | 1.756 | 1.869 | 1.741 | 1.800 | 1.744 |
| Alamat | 6.231 | 2.073 | 1.825 | 6.889 | 10.800 | 5.011 |
| Sekolah | 27 | 9.390 | 2.269 | 15.259 | 5.400 | 2.300 |
| Jml anggota | 13 | 41 | 160 | 27 | 20 | 90 |

Optimasi Jumlah *Cluster* Menggunakan Metode Elbow

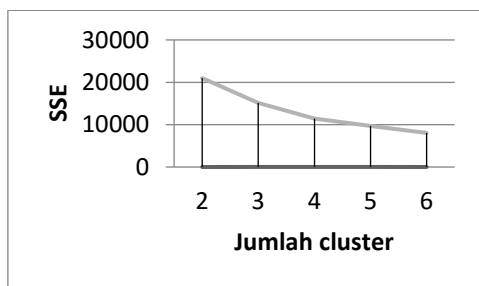
Dari pengolahan data yang sudah dilakukan dari *cluster* 0 sampai dengan *cluster* 5 diperoleh nilai SSE tiap *cluster* yang ditampilkan pada tabel dibawah ini:

- a. SSE yang diperoleh dari pengolahan menggunakan *Euclidean Distance space*

Tabel 29. Nilai SSE pada Metode *Euclidean Distance*

| Cluster | Hasil SSE | Selisih |
|---------|-----------|---------|
| 2 | 20.997 | 20.997 |
| 3 | 15.115 | 5.882 |
| 4 | 11.430 | 3.685 |
| 5 | 9.714 | 1.716 |
| 6 | 8.058 | 1.656 |

Berikut ini grafik Sum Square Error:



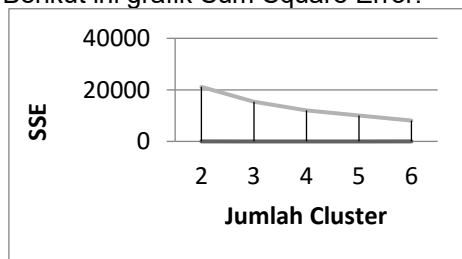
Gambar 2. Grapik Elbow Metode Euclidean Distance

- b. SSE yang diperoleh dari pengolahan menggunakan *Manhattan Distance space*

Tabel 30. Nilai SSE pada Metode *Manhattan Distance*

| cluster | hasil SSE | Selisih |
|---------|-----------|---------|
| 2 | 21156 | 21156 |
| 3 | 15398 | 5758 |
| 4 | 11980 | 3418 |
| 5 | 10074 | 1906 |
| 6 | 8078 | 1996 |

Berikut ini grafik Sum Square Error:



Gambar 3. Grapik Elbow Metode Manhattan Distance

Dari nilai SSE yang dihitung baik dari metode *Euclidean* atau *Manhattan* menyatakan bahwa jumlah *cluster* terbaik pada data siswa SMK Muhammadiyah cimanggu adalah 3 *cluster*, hal tersebut ditunjukkan dengan melihat selisih paling besar dan dari grafik dapat dilihat titik elbow berada di nilai $k=3$.

Pembahasan Hasil Penulisan

Dari hasil penulisan dapat dilihat bahwa jumlah *cluster* terbaik pada data siswa SMK Muhammadiyah Cimanggu yaitu 3 *cluster*, hal ini dapat dilihat dari hasil selisih nilai Sum Square Error (SSE), nilai selisih terbesar pada metode *Euclidean distance space* yaitu

sebesar 5.882 berada dititik *cluster* 3 dan nilai selisih terbesar pada metode *Manhattan distance space* yaitu sebesar 5.758 berada dititik *cluster* 3, Hal ini menunjukkan bahwa jumlah *cluster* optimal untuk data siswa SMK Muhammadiyah Cimanggu yaitu sebanyak tiga *cluster*.

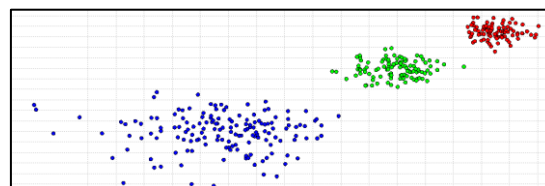
Dalam proses Perbandingan antara Metode *Euclidean* dan Metode *Manhattan*, jika dilihat dari segi jumlah iterasi, iterasi *Manhattan* lebih sedikit bila dibandingkan dengan iterasi *Euclidean* dimana nilai $k=3$ untuk metode *Manhattan* hanya 8 kali iterasi sementara *Euclidean* mencapai 19 iterasi hal ini menunjukkan bahwa dalam segi proses pengerjaan *Manhattan* lebih cepat bila dibandingkan dengan *Euclidean*. Untuk hasil Pengclusteran dengan nilai $k=3$ terdapat perbedaan.

Tabel 31. Perbandingan Avg. *within Centroid distance*

| Metode | Cluster 0 | Cluster 1 | Cluster 2 | Avg. within Centroid distance |
|------------------|-----------|-----------|-----------|-------------------------------|
| <i>Manhattan</i> | 77 | 30 | 244 | 15.398 |
| <i>Euclidean</i> | 74 | 225 | 22 | 15.115 |

jika dilihat dari Average within *Centroid distance*, *Euclidean* lebih kecil dibandingkan dengan *Manhattan* yaitu $15.115 < 15.398$, oleh karena itu *Distance measure* paling optimal untuk digunakan dalam kasus pengclusteran data siswa SMK Muhammadiyah Cimanggu adalah *Euclidean distance*.

Hasil Pengklusteran dengan nilai $k=3$ menggunakan metode *Euclidean distance* adalah *cluster* 0 sebanyak 74 data, *cluster* 1 sebanyak 225 data, dan *cluster* 2 sebanyak 22 data persebaran data dapat dilihat pada gambar 4.



Gambar 4. Persebaran Data

Penentuan strategi promosi pada SMK Muhammadiyah cimanggu dapat dianalisa dari hasil pengelompokan yang telah dilakukan penulis, adapun data yang perlu dianalisa terangkum dalam tabel 32.

Tabel 32. Data Clustering Pada Cluster 0

| Jurusan | | Alamat | | Sekolah | |
|---------|----|---------------|----|-------------------------------|----|
| TSM | 42 | Karangreja | 23 | Smpn 4 Cipari | 12 |
| AP | 22 | Cimanggu | 13 | Smp Negeri 1 Cimanggu | 10 |
| AK | 10 | Sidasari | 10 | Smp Pgri 3 Caruy | 8 |
| | | Caruy | 5 | Smpn 3 Satu Atap Cipari | 8 |
| | | Bantarmangu | 5 | Smp Muhammadiyah | 7 |
| | | Rejodadi | 4 | Smp N Pgri 3 | 7 |
| | | Panimbang | 2 | Smp Raden Fattah Cimanggu | 5 |
| | | Cilempuyang | 2 | Smp Dipo Cimanggu | 5 |
| | | Cibalung | 2 | Smp Diponegoro Cimanggu | 4 |
| | | Bantarpanjang | 2 | Smp Ma'arif Nu Cimanggu | 4 |
| | | Cilopadang | 1 | Smp Negeri 3 Satu Atap Cipari | 4 |
| | | Mekarsari | 1 | | |
| | | Kutabima | 1 | | |
| | | Kutasari | 1 | | |
| | | Bantramangu | 1 | | |
| | | Ciporos | 1 | | |

Tabel 33. Data Clustering Pada Cluster 1

| Jurusan | | Alamat | | Sekolah | |
|---------|-----|---------------|----|---------------------------|----|
| TS M | 104 | Cimanggu | 71 | Smp Muh Cimanggu | 89 |
| AK | 89 | Karangreja | 47 | Smp Pgri 3 Caruy | 74 |
| AP | 62 | Kutasari | 44 | Smp Islam Caruy | 50 |
| | | Caruy | 39 | Smpn 1 Cimanggu | 20 |
| | | Bantarpanjang | 25 | Smp Muhammadiyah Cimanggu | 19 |
| | | Bantarmangu | 16 | Smpn 4 Cipari | 3 |
| | | Panimbang | 6 | | |
| | | Sidasari | 2 | | |
| | | Cilempuyang | 2 | | |
| | | Kapuk | 1 | | |
| | | Sudimara | 1 | | |
| | | Kaangreja | 1 | | |

Tabel 34. Data Clustering Pada Cluster 2

| Jurusan | | Alamat | | Sekolah | |
|---------|---|----------------|---|-------------------------------|---|
| TSM | 9 | Karangreja | 5 | SMP Darussalam Cimanggu | 3 |
| AP | 7 | Cimanggu | 4 | SMP N Cimanggu | 2 |
| AK | 6 | Bantarpanjang | 3 | MTS Darwata Padangsari | 2 |
| | | Rejodadi | 2 | SMPN 3 Majenang | 2 |
| | | Bantarmangu | 2 | SMP Pgri 18 Cimanggu | 1 |
| | | Cilopadang | 2 | Pindahan | 1 |
| | | Cilempuyang | 1 | SMPN 01 Cimanggu | 1 |
| | | Cibalung | 1 | SMP N Satu Atap 1 Baantarsari | 1 |
| | | Cisalak | 1 | SMP Pgri 1 Caruy | 1 |
| | | Desa Citembong | 1 | SMP Negeri 1 Purwokerto | 1 |
| | | | | SMP Satu Atap Cipari | 1 |
| | | | | SMPN 1 Serang Panjang | 1 |
| | | | | SMPN 1 Cilempuyang | 1 |
| | | | | SMP Diponegoro Majenang | 1 |
| | | | | SMP PGRI | 1 |
| | | | | SMP Negeri 3 Majenang | 1 |
| | | | | SMP Negeri 1 Cimanggu | 1 |

Pada kluster pertama (*Cluster 0*) siswa didominasi oleh jurusan Teknik Sepeda Motor(TSM), kebanyakan berasal dari SMP Negeri Cipari, dan paling banyak berasal dari Kelurahan karangreja. Sedangkan siswa paling sedikit memilih Jurusan Akuntansi (AK).

Pada kluster Kedua (*Cluster 1*) siswa didominasi oleh jurusan Teknik Sepeda Motor(TSM), kebanyakan berasal dari SMP Muhammadiyah Cimanggu, dan paling banyak berasal dari Kelurahan cimanggu. Sedangkan siswa paling sedikit memilih Jurusan Administrasi Perkantoran (AP).

Pada kluster ketiga (*Cluster 2*) siswa didominasi oleh jurusan Teknik Sepeda Motor(TSM), kebanyakan berasal dari SMP Darussalam Cimanggu, dan paling banyak berasal dari Kelurahan karangreja. Sedangkan siswa paling sedikit memilih Jurusan Akuntansi (AK).

Secara keseluruhan kluster diperoleh informasi bahwa kebanyakan pilihan siswa pada Jurusan Teknik Sepeda Motor yang

beralamatkan dari Kelurahan Cimanggu dimana kebanyakan siswa berasal dari SMP Muhammadiyah Cimanggu, Sementara Jurusan dengan minat paling sedikit yaitu jurusan akuntansi, dari data pula dapat dilihat untuk kelurahan sudimara, cisalak, cileumpuyang, citembong dan kapuk sangat minim peminat.

Informasi yang diperoleh dari masing-masing kluster secara detail dapat digunakan untuk menentukan strategi promosi SMK Muhammadiyah Cimanggu dengan mempertimbangkan nilai-nilai minimal yang perlu diambil tindakan dalam strategi promosi selanjutnya.

Penentuan strategi promosi SMK Muhammadiyah Cimanggu dapat dilakukan dengan beberapa strategi berdasarkan hasil *cluster* yang masih menunjukkan nilai minimal. Dengan memetakan variabel-variabel yang terdapat nilai minimal, memungkinkan peningkatan variabel yang merata.

4. Kesimpulan

Berdasarkan hasil analisis terhadap pengelompokan data siswa, dapat disimpulkan bahwa Penentuan jumlah *cluster* terbaik dengan metode elbow dapat menghasilkan jumlah *cluster* K yang sama pada metode *Euclidean distance space* dan *Manhattan distance space*, Hasil penentuan jumlah *cluster* terbaik dengan metode elbow akan dijadikan *default* untuk proses karakteristik berdasarkan studi kasus yang dilakukan dimana jumlah *cluster* terbaik yaitu sebanyak 3 *cluster* yaitu Dalam proses pengerjaan secara manual metode *Manhattan* lebih cepat bila dibandingkan dengan *Euclidean* dimana jumlah iterasi pada *Manhattan* selalu lebih sedikit dari *Euclidean*, *Distance measure* paling optimal untuk digunakan dalam kasus pengklasteran data siswa SMK Muhammadiyah Cimanggu adalah *Euclidean distance*. Dalam penentuan strategi promosi SMK Muhammadiyah Cimanggu dapat dilakukan dengan mempertimbangkan nilai-nilai minimal yang perlu diambil tindakan dalam strategi promosi, Pendekatan ini memungkinkan memiliki peluang besar untuk meningkatkan jumlah pendaftar siswa baru.

Referensi

- Anggara, M., Sujiani, H., & Helfi, N. (2016). Pemilihan Distance Measure Pada K-Means Clustering Untuk Pengelompokan Member Di Alvaro Fitness. *Jurnal Sistem Dan Teknologi Informasi*, 1(1), 1–6.
- Burhanuddin, A., Utami, E., & Pramono, E. (2017). Perbandingan Metode Single Linkage dan Fuzzy C Means Untuk Pengelompokan Trafik Internet, *XII*, 1–6.
- Chasanah, T. T., & Widiyono. (2017). Penentuan Strategi Promosi Penerimaan Mahasiswa Baru dengan Algoritma Clustering K-Means. *IC-Tech*, *XII*(1), 39–44.
- Eka Merliana, N. P., Ernawati, & Santoso, A. J. (2015). Analisa Penentuan Jumlah Cluster Terbaik Pada Metode K-Means. *Prosiding Seminar Nasional Multi Disiplin Ilmu&Call for Papers Unisbank (Sendi_U)*, 978–979.
- Hedynata, M. L., & E.D.Radianto, W. (2016). Strategi Promosi Dalam Meningkatkan Penjualan Luscious Chocolate Potato Snack. *Jurnal Manajemen*, 1(1), 87–96. Retrieved from
- Hendrawan, B. (2015). Artikel Skripsi Unersitas Nusantara PGRI Kediri. *Implementasi Sms Getaway Untuk Layanan Informasi Absensi Siwa*, 01(11), 1–16. <https://doi.org/10.103.03.0022>
- Ong, J. O. (2013). Implementasi Algoritma K-means clustering untuk menentukan strategi marketing president unersity. *Jurnal Ilmiah Teknik Industri*, vol.12, no(juni), 10–20.
- Sulastri, H., & Gufroni, A. I. (2017). Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia. *Jurnal Teknologi Dan Sistem Informasi*, 3(2), 299.
- Yoga Religia. (2016). Metode Manhattan, Euclidean Dan Chebyshev Pada Algoritma K-Means Untuk Pengelompokan Status Desa.
- Yudi Agusta. (2007). NIMH report presents latest data on inpatient psychiatric services for patients 65 and over. *News & Notes*, 3(11), 47–60.