PENERAPAN ALGORITMA K-MEANS *CLUSTERING* ANALYSIS PADA PENYAKIT MENULAR MANUSIA (STUDI KASUS KABUPATEN MAJALENGKA)

Ade Bastian, **Harun Sujadi**, **dan Gigin Febrianto**

Program Studi Teknik Informatika, Universitas Majalengka, Jl. Universitas Majalengka No.1, Majalengka, 45452, Indonesia

adb@ft.unma.ac.id, harunsujadi@gmail.com, gien_feb@gmail.com

Abstract

Health is a valuable thing for humans because anyone can experience health problems, as well as in humans are very susceptible to various diseases but the cause we do not realize. The K-means algorithm is not affected by the order of the objects used, it is proved when the author tries to randomly determine the starting point of the *cluster* center of one of the objects at the beginning of the calculation. The number of *cluster* members generated amounts to the same when using other objects as the starting point of the *cluster* center. However, this only affects the number of iterations performed. Object grouping (object *clustering*) is one of the processes of object mining that aims to partition an existing object into one or more *cluster* objects based on its characteristics. This study examines how the use of K-means *Cluster* Analysis method in the case study of human contagious diseases on an object. This study examines the K-means *Cluster* Analysis method in infectious diseases in humans based on the set of variables established per sub-district of each Puskesmas in which there are 32 Puskesmas offices in Majalengka District.

Keywords: *Algoritm,* K-means*, Clustering, Human Infectious Diseases,Puskesmas*

Abstrak

Kesehatan merupakan hal yang berharga bagi manusia karena siapa saja dapat mengalami gangguan kesehatan, begitu pula pada manusia yang sangat rentan terhadap berbagai macam penyakit namun penyebabnya tidak kita sadari. Algoritma K-means tidak terpengaruh terhadap urutan objek yang digunakan, hal ini dibuktikan ketika penulis mencoba menentukan secara acak titik awal pusat *cluster* dari salah satu objek pada permulaan perhitungan. Jumlah keanggotaan *cluster* yang dihasilkan berjumlah sama ketika menggunakan objek yang lain sebagai titik awal pusat *cluster* tersebut. Namun, hal ini hanya berpengaruh pada jumlah iterasi yang dilakukan. Pengelompokan objek (objek *clustering*) adalah salah satu proses dari objek mining yang bertujuan untuk mempartisi objek yang ada kedalam satu atau lebih *cluster* objek berdasarkan karakteristiknya. Penelitian ini mengkaji bagaimana penggunaan Algoritma K-means *Cluster Analysis* dalam studi kasus penyakit menular manusia pada suatu objek. Penelitian ini mengkaji metode K-means *Cluster Analysis* dalam penyakit menular pada manusia berdasarkan set variabel yang dibentuk per kecamatan tiap Puskesmas yang jumlahnya ada 32 Kantor Puskesmas di Kabupaten Majalengka.

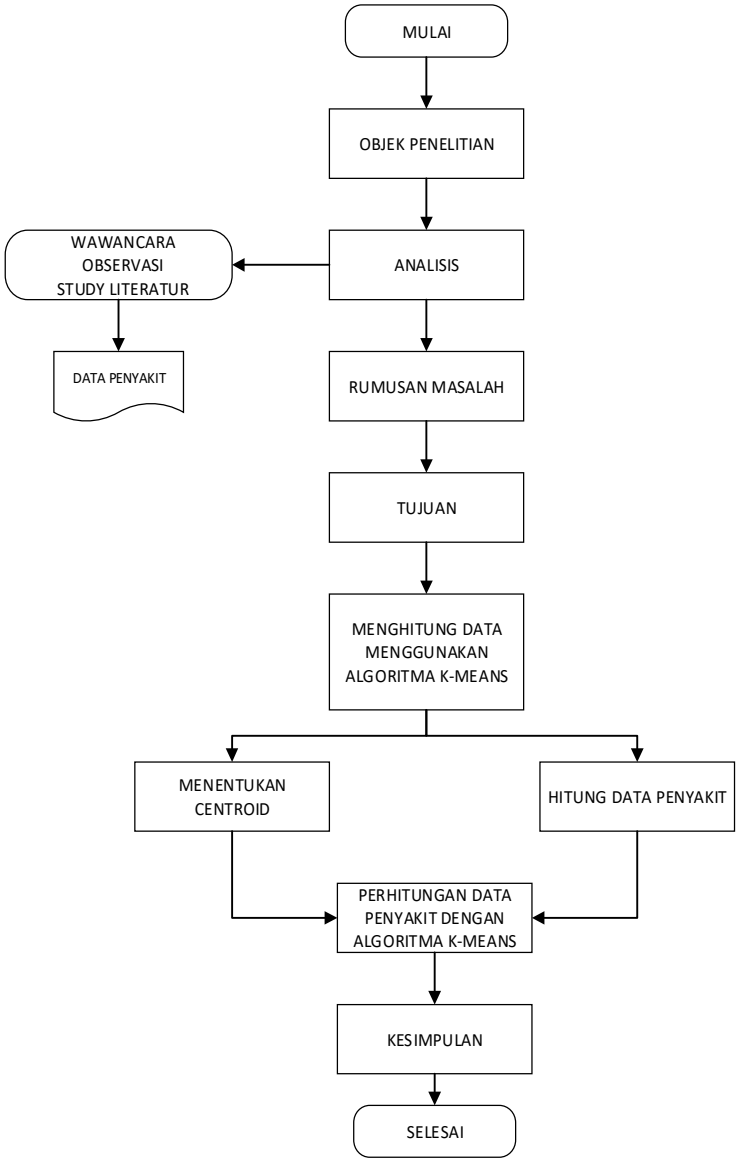
Kata Kunci: Algoritma, K-means*, Clustering, Penyakit Menular Manusia, Puskesmas*

1. Pendahuluan

Kemajuan teknologi informasi sudah semakin berkembang pesat di segala bidang kehidupan. Banyak sekali data yang dihasilkan oleh teknologi informasi yang semakin canggih, mulai dari bidang industri, ekonomi, ilmu dan teknologi serta berbagai bidang kehidupan lainnya. Penerapan teknologi informasi dalam dunia kesehatan juga dapat menghasilkan data yang berlimpah mengenai penyakit menular manusia [1].

Kesehatan merupakan hal yang berharga bagi manusia karena siapa saja dapat mengalami gangguan kesehatan, begitu pula pada manusia yang sangat rentan terhadap berbagai macam penyakit tetapi penyebabnya tidak kita sadari. Hambatan-hambatan yang menyebabkan sulitnya melakukan konsultasi penyakit oleh dokter sekarang ini dapat diatasi dengan adanya program komputer. Dalam hal ini, data informasi dapat membantu pemecahan masalah terhadap penyakit-penyakit dengan diberikan nasihat kepada pembaca dan menemukan solusi terhadap berbagai permasalahan yang spesifik.

Analisis *cluster* merupakan teknik multivariat yang mempunyai tujuan utama mengelompokan penyakit menular manusia berdasarkan yang frekuensi kejadian. Pada penelitian ini, analisis *cluster* dilakukan di daerah Majalengka. Analisis *Clus-*

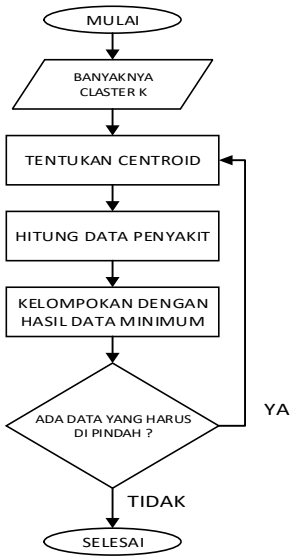


Gambar 1. Kerangka Penelitian

ter penyakit menular pada manusia dilakukan sehingga setiap penyakit yang paling banyak kesamaannya dengan objek lain akan berada dalam *cluster* yang sama. *Cluster-cluster* yang terbentuk memiliki *homogenitas* internal dan heterogenitas eksternal yang tinggi.

Pada penelitian ini, diimplementasikan algoritma *clustering* K-means. Alasan penggunaan algoritma K-means di antaranya ialah karena algoritma ini memiliki ketelitian yang cukup tinggi terhadap ukuran objek, sehingga *algoritma* ini relatif lebih terukur dan efisien untuk pengolahan objek dalam jumlah besar. Selain itu *algoritma* K-means ini tidak terpengaruh oleh urutan objek.

Tujuan perancangan program ini bukan untuk menggantikan peran manusia, tetapi untuk mensubtitusikan pengetahuan manusia kedalam bentuk sistem agar dapat di gunakan oleh orang banyak. Penelitian ini mendiskusikan dan memperlihatkan metode pengelompokkan pada data penyakit yang dimulai dengan membangun data *clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan (*unsupervised*). Ada dua jenis data *clustering* yang sering digunakan dalam proses pengelompokan data yaitu *hierarchical* (hirarki) *data clustering* dan *non-hierarchical* (non hirarki) *data clustering.* K-means merupakan salah satu metode data *clustering* non hirarki yang berusaha



Gambar 2. Flowchart Algoritma Metode K-Means

mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster/*kelompok. Metode ini mempartisi data ke dalam *cluster/*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. Adapun tujuan dari data *clustering* ini adalah untuk meminimalisasi *objective function* yang ditentukan pada saat proses *clustering*, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu *cluster* dan memaksimalisasikan variasi antar *cluster* [2].

Permasalahan yang dikaji dalam penelitian ini adalah bagaimana penggunaan metode K-means *Cluster Analysis* dalam penyakit menular manusia pada suatu objek. Tujuan yang ingin penulis capai adalah mengkaji metode K-means *Cluster Analysis* dalam penyakit menular pada manusia berdasarkan set variabel yang dibentuk per puskesmas di tiap kecamatan yang jumlahnya ada 32 kantor Puskesmas di Kabupaten Majalengka. Terdapat beberapa penyakit menular yang berjangkit di Kabupaten Majalengka. Enam di antaranya adalah penyakit Diare, ISPA, Kusta, Malaria, Tuberkulosa, Penyakit Menular Seks. Seluruh data diambil dari sejumlah Puskesmas di Kabupaten Majalengka.

2. Metode

Alur penelitian mengikuti kerangka penelitian pada Gambar 1. Penelitian dilakukan di Dinas Kesehatan Kabupaten Majalengka. Penelitian diawali dengan analisis yang meliputi wawancara ke setiap karyawan atau pegawai di Dinas Kesehatan, lalu

TABEL 1.
TITIK PUSAT AWAL DARI TIAP CLASTER ITERASI KE-1

	Diare	Ispa	Kusta	Tuberkulosis	Malaria	Penyakit seks
C1	821	1679	235	310	159	31
C2	570	1575	285	969	97	27
C3	1057	2363	261	483	171	82
C4	927	1547	307	469	480	36
C5	732	472	607	472	204	29
C6	843	8	334	162	63	24

dilanjutkan dengan penelitian terhadap data penyakit manusia, kemudian menentukan penyakit mana saja yang akan diambil sampelnya. Setelah itu, dilakukan *study literature* tentang teori-teori untuk mengolah data tersebut. Analisis ini menghasilkan data pe-nyakit menular manusia.

Penelitian dilanjutkan dengan merumuskan masalah agar penelitian ini terarah dan mempunyai masalah yang melahirkan sebuah tujuan. Tujuannya yaitu proses pengelompokan *data mining* yang menerapkan *algoritma* K-means *clustering* pada penyakit menular manusia di Kabupaten Majalengka untuk data tahun 2014. Selanjutnya, hasil pengelompokan dengan algoritma K-means dianalisis untuk menentukan penyakit mana yang banyak diderita.

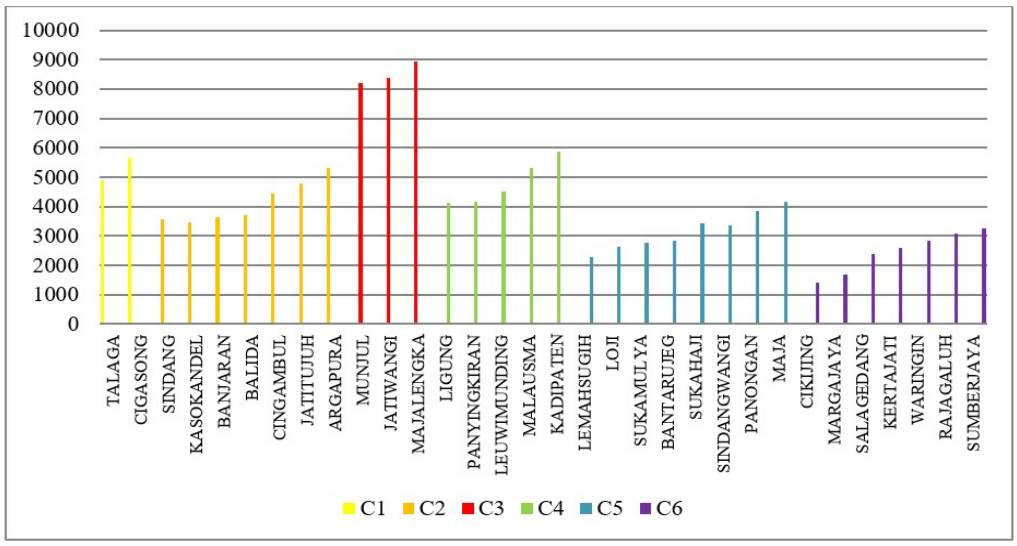
Data mining

Istilah *data mining* memiliki beberapa pandangan, seperti *knowledge discover* ataupun *pattern recognition.* Kedua istilah tersebut sebenernya memiliki ketepatannya masing-masing, istilah *knowledge discovery* atau penemuan pengetahuan tepat karna digunakan tujuan utama dari *data mining* memang untuk mendapat pengetahuan yang masih tersembunyi di dalam bongkahan data [3,4]. Istilah *pattern recognition* atau pengenalan pola pun tetap untuk digunakan karena pengetahuan yang hendak digali memang berbentuk pola-pola yang juga masih perlu digali dari dalam bongkahan data yang tengah dihadapi [4].

Banyak definisi bagi istilah ini dan belum ada yang dibakukan atau disepakati semua pihak. Namun demikian, istilah ini memiliki hakikat (*notion*) sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki, kegiatan inilah yang menjadi garapan atau perhatian utama dari disiplin ilmu *data mining* [5].

Fungsi-Fungsi Dari *Data mining*

Dalam rangka menemukan, menggali, atau menambang pengetahuan, terdapat enam fungsi dalam *data mining*, yaitu [2]: 1) Fungsi deskripsi (*description*), 2) Fungsi estimasi (*estimation*), 3) Fung-



Gambar 3. Grafik Puskesmas Tiap *Cluster*

si prediksi (*prediction*), 4) Fungsi klasifikasi (*classification*), 5) Fungsi pengelompokan (*classification*), dan 6) Fungsi asosiasi (*association*).

Mengacu kepada keenam fungsi *data mining* tersebut dapat dipilih menjadi [2]: 1) Fungsi minor atau fungsi tambahan, yang meliputi ketiga fungsi yang pertama, yaitu deskripsi. Estimasi, dan prediksi; dan 2) Fungsi mayor atau fungsi utama, yang meliputi ketiga fungsi berikut, yaitu klasifikasi, pengelompokan, dan asosiasi.

Clustering

Pada dasarnya *clustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteriktik (*similarity*) antara satu data dengan data yang lain. *Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan (*unsupervised*), maksudnya metode ini diterapkan tanpa adanya latihan (*training*)dan tanpa ada guru serta tidak memerlukan target *output*. Dalam *data mining* ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering* [6].

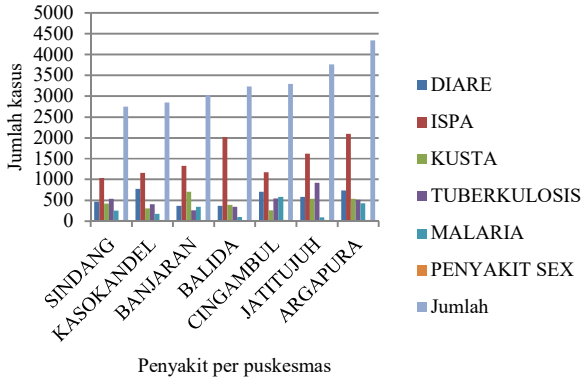
Hierarchical *clustering* adalah suatu metode pengelompokan data yang dimulai dengan mengelompokkan dua atau lebih objek yang memiliki kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang memiliki kedekatan kedua. Demikian seterusnya sehingga *cluster* akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya akan membentuk sebuah *cluster*. Dendogram biasanya digunakan untuk membantu memperjelas proses hierarki tersebut [6].

Berbeda dengan metode hierarchical *clustering*, metode non-hierarchical *clustering* justru dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan (dua *cluster*, tiga *cluster*, atau lain sebagainya). Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan tanpa mengikuti proses hierarki. Metode ini biasa disebut dengan K-means *Clustering* [7].

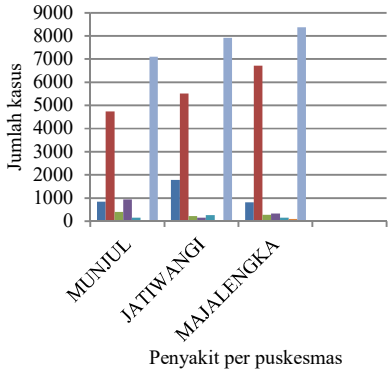
Flowchart Algoritma Metode K-means

Metode K-means *clustering* merupakan metode *clustering* yang dikenalkan oleh [8]. Metode K-means adalah metode yang terkenal cepat dan simpel [9]. K-means *clustering* merupakan salah satu metode data *clustering* non-hirarki yang mengelompokan data dalam bentuk satu atau lebih *cluster*/kelompok. Data-data yang memiliki karakteristik yang sama dikelompokan dalam satu *cluster*/kelompok dan data yang memiliki karakteristik yang berbeda dikelompokan dengan *cluster*/kelompok yang lain sehingga data yang berada dalam satu *cluster*/kelompok memiliki tingkat variasi yang kecil [7].

Langkah-langkah melakukan *clustering* dengan metode K-means adalah sebagai berikut [7]: 1) Pilih jumlah *cluster* k; 2) Inisialisasi ke pusat *cluster* ini bisa dilakukan dengan berbagai cara. Cara yang paling sering dilakukan adalah dengan random atau acak. Pusat-pusat *cluster* diberiduberi nilai awal dengan angka-angka random; 3) Alokasikan semua data/objek ke *cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu



Gambar 4. Cluster 2



Gambar 5. Cluster 3

data ke *cluster* tertentu ditentukan jarak antara data dengan pusat *cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster*. Jarak paling antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk dalam *cluster* mana. Untuk menghiutng jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai persamaan(1) berikut.

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \ \dots \ + (X_{ki} - X_{kj})^2} \quad (1)$$

dimana:
$D(i,j)$ = Jarak data keike pusat *cluster j*
X_{ki} = Data ke *i* pada atribut data ke *k*
X_{kj} = Titik pusat ke *j* pada atribut ke *k*

Jarak pusat *cluster* dihitung kembali dengan keanggotaan *cluster* yang sekarang. Pusat *cluster* adalah rata-rata dari semua data/objek dalam *cluster* tertentu. Jika dikehendaki bisa juga menggunakan median dari *cluster* tersebut. Jadi rata-rata (*mean*) bukan satu-satunya ukuran yang bisa dipakai. Setiap objek kemudian ditugaskan kembali memakai pusat *cluster* yang baru. Jika pusat *cluster* tidak berubah lagi maka proses *clustering* selesai. Atau, kembali ke langkah nomor 3 sampai pusat *cluster* tidak berubah lagi.

Berdasarkan gambar 2, langkah-langkah yang dilakukan oleh algoritma metode K-means adalah sebagai berikut [6]: 1) Mulai; 2) Menentukan banyaknya *cluster* yaitu dari data penyakit menular; 3) Pengesetan nilai awal titik tengah/*centoid*. Langkah ketiga, menentukan pusat *cluster* secara acak pada data awal; 4) Menghitung data penyakit ke *centroid* dengan menggunakan rumus jarak *Euclid*; 5) Melakukan *clustering* data dengan memasukkan setiap obyek ke dalam *cluster* (grup) berdasarkan jarak minimumnya; 6) Jika ada data yang harus dipindah, maka langkah selanjutnya adalah menghitung pusat *cluster* baru. Pusat *cluster* yang baru ditentukan berdasarkan pengelompokan anggota masing-masing *cluster* baru. Pusat *cluster* baru untuk *cluster* yang pertama dihitung berdasarkan rata-rata koordinat. Pengulangan dihentikan sampai hasil perhitungan menunjukkan adanya angka pusat *cluster* yang sama; 7) Selesai.

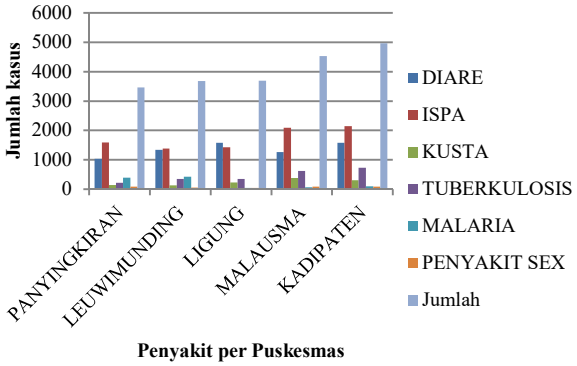
3. Hasil dan Analisis

Setelah semua data penyakit menular dari Dinas Kesehatan pada tahun 2014 terkumpul, maka data-data tersebut telah dapat dikelompokan dengan menggunakan algoritma K-means *Clustering*. Cara penghitungannya dengan mengukur *Euclidean distance*-nya. Pengukuran dilakukan terhadap dua titik dalam satu, dua dan tiga dimensi secara berurutan. Untuk dapat melakukan pengelompokan data-data tersebut menjadi beberapa *cluster* perlu dilakukan beberapa langkah, yaitu: 1) Menentukan jumlah *cluster*, dalam penelitian ini data-data yang ada akan dikelompokkan mejadi 6 *cluster* (Gambar 3); serta 2) Menentukan titik pusat awal dari setiap *cluster*. Dalam penelitian ini, titik pusat awal ditentukan dengan menghitung dari rata-rata data terkecil dan terbanyak dan didapat titik pusat dari setiap *cluster* dapat dilihat pada Tabel 1.

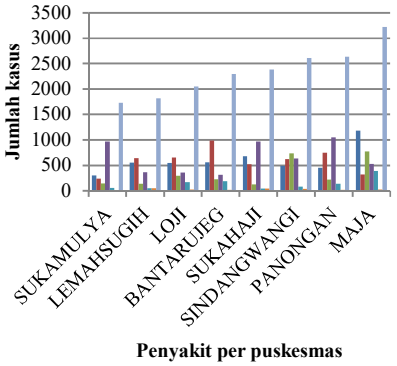
Pada tahap ini perlu dihitung jarak tiap data ke tiap pusat *cluster*. Jarak paling dekat antara satu data dengan satu *cluster* tertentu akan menentukan suatu data masuk dalam *cluster* mana. Untuk menghitung jarak semua data ke setiap titik pusat *cluster* dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai persamaan(2) berikut:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \ \dots + \ (X_{ki} - X_{kj})^2} \quad (2)$$

dimana:
$D(i,j)$ = jarak data ke i ke pusat *cluster j*



Gambar 6. Cluster 4



Gambar 7. Cluster 5

X_{ki} = data ke i pada atribut data ke k
X_{kj} = titik pusat ke j pada atribut ke k

Pengulangan dihentikan karena hasil perhitungan menunjukkan adanya angka pusat *cluster* yang sama pada iterasi ke-5 dan ke-6.

Seperti yang ditunjukkan oleh grafik pada Gambar 4, dari hasil *cluster* 1, terdiri dari 9416 jiwa penderita penyakit menular yang berasal dari Puskesmas Cigasong dan Talaga, terlihat bahwa penyakit menular pada *cluster* 1 didominasi oleh penyakit ISPA dan Diare.

Sementara itu, dari hasil *cluster* 2 (ditunjukkan pada Gambar 5), terdiri dari 23245 jiwa penderita penyakit menular yang berasal dari Puskesmas Cingambul, Banjaran, Argapura, Sindang, Balida, Kasokandel dan Jatitujuh terlihat bahwa penyakit menular pada *cluster* 2 didominasi oleh penyakit ISPA dan Diare. Sedangkan, berdasarkan Puskesmas didominasi oleh penyakit yang berasal dari Puskesmas Argapura dan Balida.

Gambar 6 menunjukkan hasil *cluster* 3. *Cluster* ini terdiri dari 23415 jiwa penderita penya-kit menular yang berasal dari Puskesmas Majalengka, Munjul dan Jatiwangi, dapat dilihat bahwa penyakit menular pada *cluster* 3 didominasi oleh penyakit ISPA dan Diare. Sedangkan, berdasarkan Puskesmas didominasi oleh penyakit yang berasal dari Puskesmas Majalengka dan Jatiwangi.

Hasil *cluster* 4 seperti diberikan pada Gambar 7, terdiri dari 20334 jiwa penderita penyakit menular yang berasal dari Puskesmas Malausma, Leuwimunding, Panyingkiran, Kadipaten dan Ligung, terlihat bahwa penyakit menular pada *cluster* 4 didominasi oleh penyakit ISPA dan Diare. Dan sedangkan, berdasarkan Puskesmas didominasi oleh penyakit yang berasal dari puskesmas Kadipaten dan Malausma.

Hasil *cluster* 5 (Gambar 8), terdiri dari 16384 jiwa penderita penyakit menular yang berasal dari Puskesmas Lemahsugih, Bantarujeg, Maja, Sukahaji, Sindangwangi, Loji, Sukamulya dan Panongan, terlihat bahwa penyakit menular pada *cluster* 5 didominasi oleh penyakit Tuberkulosis dan Diare. Sedangkan, berdasarkan Puskesmas didominasi oleh penyakit yang berasal dari Puskesmas Panongan dan Maja.

Cluster 6, terdiri dari 13143 jiwa penderita penyakit menular yang berasal dari Puskesmas Margajaya, Cikijing, Salagedang, Rajagaluh, Waringin, Kertajati dan Sumberjaya. Pada Gambar 9 terlihat bahwa penyakit menular pada *cluster* 6 didominasi oleh penyakit Diare dan Tuberkulosis. Sedangkan, berdasarkan Puskesmas didominasi oleh penyakit yang berasal dari Puskesmas Kertajati dan Sumberjaya.

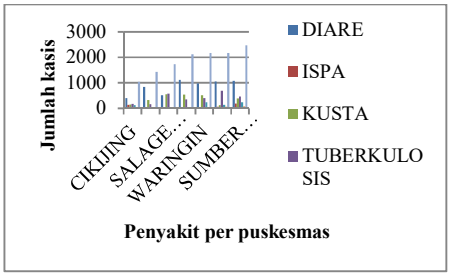
4. Kesimpulan

Dari hasil penelitian yang telah dilakukan dapat disimpulkan bahwa K-means merupakan salah satu metode data *clustering* non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster* atau kelompok. Kelebihan Algoritma K-means diantaranya adalah mampu mengelompokkan objek besar dan pencilan obyek dengan sangat cepat sehingga mempercepat proses pengelompokan.

Kekurangan Algoritma K-means yaitu. sangat sensitive pada pembangkitan titik pusat awal secara random, hasil pengelompokan bersifat tidak unik (selalu berubah-ubah) dan proses pengerjaannya cepat tetapi keakuratannya tidak dijamin.

Dari hasil penelitian dapat ditarik kesimpulan bahwa hasil dari metode Algoritma K-means *clustering data mining* dapat digunakan untuk metode pengendalian persediaan pada Puskesmas Pandanaran, sehingga apabila akan dilakukan pengadaan persediaan obat pada tahun 2014, petugas dapat melihat daftar Puskesmas terbanyak yang menderita penyakit menular.

Dari data yang diolah berdasarkan jenis ba-



Gambar 8. Cluster 6

rang, diinputkan sampel data sebanyak 32 data dengan. Jumlah yang diperoleh 6 kelompok data telah ter*cluster*.

Referensi

[1] Cushing, B.E., Graham Jr, L.E., Palmrose, Z.V., Roussey, R.S. and Solomon, I., 1995. Risk orientation. Auditing, Practice, Research, and Education A Productive Collaboration.

[2] Larose, D.T., 2005. Introduction to data mining (pp. 1-26). John Wiley & Sons, Inc..

[3] Berry, M.W. and Browne, M., 2006. Lecture notes in data mining. World Scientific.

[4] Susanto, S. and Suryadi, D., 2010. Pengantar data mining: mengagali pengetahuan dari bongkahan data.

[5] Luthfi, K. and Taufiq, E., 2009. Algoritma Data Mining. Yogyakarta: Andi.

[6] Santosa, B., Conway, T. and Trafalis, T., 2007. A hybrid knowledge based-clustering multi-class SVM approach for genes expression analysis. In Data Mining in Biomedicine (pp. 261-274). Springer, Boston, MA.

[7] Agusta, Y., 2007. K-Means Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika, 3(1), pp.47-60.

[8] Lloyd, S., 1982. Least squares quantization in PCM. IEEE transactions on information theory, 28(2), pp.129-137.

[9] Arthur, D. and Vassilvitskii, S., 2006, June. How slow is the k-means method?. In Proceedings of the twenty-second annual symposium on Computational geometry (pp. 144-153). ACM..