SYSTEM FOR DETERMINING PUBLIC HEALTH LEVEL USING THE AGGLOMERATIVE HIERARCHICAL CLUSTERING METHOD

Suhirman[1], Hero Wintolo[2]
[1]Program Studi Informatika, Fakultas Teknologi Informasi dan Elektro
Universitas Teknologi Yogykarta
Jl. Ringroad Utara Jombor Sleman Yogyakarta
[2]Program Studi Teknik Informatika
Sekolah Tinggi Teknologi Adisutjipto
Jl. Jjanti Blok R, Lanud Adisutjipto, Yogyakarta
E-mail : [1]suhirman@uty.ac.id, [2]herowintolo@stta.ac.id

Abstract

Regions having higher level of welfare do not always have better indicator values than other regions having lower level of welfare. The problem is the lack of information related to the indicator values needed to determine the health level. Therefore, clustering using health data becomes necessary. Data were clustered to see the maximum or the minimum level of similarity. The clustered data were based on the similarity of four morality indicator values of the regional health level. Morality indicator values used in this research are infant mortality rate, child mortality rate, maternal mortality rate, and rough birth rate. The method used is Agglomerative Hierarchical Clustering (AHC) - Complete Linkage. Data were calculated using Euclidean Distance Equation, then Complete Linkage. Four clustered data were grouped into two clusters, healthy and/or unhealthy. The result, combining from all clusters into two large clusters to see healthy and unhealthy results.

Keywords: *Health Level, Health Indicators, Agglomerative Hierarchical Clustering, Cluster.*

1. Latar Belakang

Kesehatan adalah keadaan sejahtera baik keadaan fisik, rohani dan sosial yang memungkinkan seseorang untuk dapat melakukan segala kegiatan. Menurut Organisasi Kesehatan Dunia (WHO) tahun 1948 menyebutkan bahwa pengertian kesehatan adalah sebagai "suatu keadaan fisik, mental, dan sosial kesejahteraan dan bukan hanya ketiadaan penyakit atau kelemahan".

Salah satu indikator tingkat keberhasilan pembangunan negara dapat dilihat dari tingkat capaian negara tersebut dalam memberikan jaminan bidang kesehatan. Penetapkan sejumlah indikator sebagai tolak ukur kemajuan pembangunan kesehatan di berbagai level dari tingkat pusat sampai daerah. Setiap tahun data kesehatan penduduk dikumpulkan dan diolah sehingga dihasilkan ranking suatu daerah atau kota sehat. Hasil olahan dan perangkingan ini merupakan acuan penting, seperti bagi pemerintah daerah untuk membuat program intervensi yang lebih tepat sasaran. Hal ini sebagai bahan advokasi agar Pemerintah daerah lebih terpacu untuk menaikkan ranking kesehatan daerah demi membangun daerah atau kota yang sehat.

Metode pengolahakan data indikator kesehatan yang selama ini diterapkan masih berbasis pada teknik statistik dasar, yaitu perhitungan yang didasarkan pada hasil rata-rata seluruh indikator atau didasarkan pada distribusi data. Ternyata output yang dihasilkan dari metode ini

masih memiliki sejumlah permasalahan terutama dalam hal konsistensi data. Dalam hal ini, daerah yang memiliki ranking lebih tinggi belum tentu memiliki semua nilai indikator yang lebih baik dibandingkan dengan daerah di ranking yang lebih rendah. Permasalahan lain terletak pada minimnya informasi tentang hubungan antar indikator. Seperti pada derajat kesehatan daerah belum ada analisis yang dilakukan untuk mengukur, misal, pengaruh status gizi masyarakat terhadap kejadian mortalitas (kematian) maupun morbiditas (kesakitan).

Clustering adalah salah satu alat dasar yang tersedia, untuk memahami struktur data[1]. Proses pengelompokan satu objek fisik atau abstrak ke dalam kelas-kelas objek serupa dikenal sebagai clustering. Teknik pengelompokan berperan penting dalam machine learning, data mining, pengambilan informasi, pemasaran, diagnostik medis, dan pengenalan pola. Clustering sering disebut tugas belajar tanpa pengawasan karena tidak ada kelas yang menunjukkan nilai dari clustering sebelumnya yang diberikan dari sampel data. Secara umum dari pengelompokan bisa menjadi "proses pengorganisasian objek ke dalam kelompok yang anggotanya mirip dalam beberapa hal". Oleh karena itu, cluster adalah kumpulan objek data yang mirip satu sama lain dalam cluster yang sama dan berbeda dengan objek di cluster lain.

Clustering merupakan bidang penelitian data mining yang dinamis. Banyak algoritma pengelompokan terus dikembangkan. Algoritma clustering tergantung pada jenis data yang tersedia dan tujuan dari suatu aplikasi [1]. Teknik clustering dapat dikategorikan menjadi:

a. Hierarchical
b. Partitional
c. Density Based
d. Grid Based

Makalah ini membahas tentang pemanfaatan *algoritma hierarchial clustering* untuk mengklaster atau mengelompokkan data derajat kesehatan berdasarkan kemiripan karakteristik daerah yang ditinjau dari nilai empat data ukuran indikator mortalitas derajat kesehatan, yaitu angka kelahiran kasar, angka kematian bayi, angka kematian balita, dan angka kematian ibu. Secara umum, output penelitian diharapkan berupa hasil klasterisasi dapat digunakan sebagai sebuah acuan untuk menggambarkan distribusi pengelompokkan data berdasarkan kondisi profil kesehatan suatu daerah.

2. Metodologi Penelitian

Setiap hari di dunia ini terjadi transaksi dan pengumpulan data dalam jumlah yang sangat besar. Analisis data ini merupakan proses yang sangat penting. Data mining dapat memenuhi kebutuhan ini. Cara yang disediakan untuk menemukan pengetahuan dari data yang sangat besar. Proses ini dapat dilihat sebagai hasil dari evolusi alami informasi.

Data mining adalah subjek yang benar-benar interdisipliner. Data mining dapat didefinisikan dalam berbagai teknologi dengan cara nyata [2]. Mining adalah istilah yang secara jelas mencirikan proses menemukan sekumpulan kecil nugget berharga dari banyak data. Selain itu, banyak istilah lain memiliki arti yang sama seperti data mining misalnya, pengetahuan tentang pengukuhan, ekstraksi pengetahuan, analisis / pola data. Penemuan pengetahuan dari data, sementara yang lain melihat data mining sebagai langkah penting dalam proses penemuan pengetahuan.

Secara umum, data mining dapat diterapkan untuk semua jenis data. Data mining banyak digunakan dalam aplikasi database, data warehouse, dan data transaksional. Data mining juga

dapat diterapkan ke bentuk lain dari data misalnya, aliran data, urutan data, grafik atau jaringan data, data spasial, data teks, dan data multimedia, dan semua jenis data pesanan yang muncul.

Analisis Cluster adalah metode analisis data penting yang digunakan untuk mengelompokkan kumpulan data dengan karakteristik serupa. Ini telah digunakan di banyak bidang seperti pemrosesan data [3], pengolahan data transaksional [4], dukungan keputusan [5], dan pemrosesan sinyal radar [6]. Sebagian besar algoritma pengelompokan berfokus pada data yang dicirikan oleh atribut numerik, di mana jarak geometri alami antara objek yang akan dikelompokkan tersedia. Banyak perhatian telah diberikan pada pengelompokan data kategori [7][8][9], di mana objek data terdiri dari atribut non-numerik. Data kategori kategorisasi menyajikan tantangan khusus karena tidak adanya properti data yang diwarisi secara inheren.

Dengan menggunakan clustering, daerah padat dapat diidentifikasi, pola distribusi dan keseluruhan hubungan yang menarik dapat dilakukan antara atribut data. Upaya data mining berfokus pada penemuan metode untuk clusteringdalam basis data yang besar secara efektif dan efisien. Jumlah pendekatan clustering menyulitkan untuk menentukan kualitas ukuran universal. Namun, beberapa hal yang perlu diperhatikan, adalah parameter input yang tidak menyulitkan pengguna. Hasil cluste dapat dianalisis, dan skalabilitasnya dapat meningkatkan ukuran dan dimensi catatan kumpulan data. Pekerjaan sebelumnya dalam menginduksi dimensi clustering telah difokuskan untuk menghasilkan banyak kumpulan data, dan secara luas dapat dibagi menjadi dua kategori:

Banyak metode clustering telah dikembangkan, masing-masing menggunakan prinsip induksi yang berbeda. Metode clustering dibagi menjadi dua kelompok, yaitu : metode hierarkis dan partisi [10]. Sedangkan pengkategorian metode dapat dibagi menjadi tiga [2] yaitu : metode berbasis kerapatan, metode berbasis model, dan metode berbasis grid.

a. Hierarchical clustering: Metode ini membangun klaster dengan secara rekursif mempartisi instance dalam mode top-down atau bottom-up. Metode-metode ini dapat dibagi sebagai berikut:
   - Aglomerative hierarchical clustering - Setiap objek awalnya merupakan kelompok tersendiri. Kemudian cluster secara berturut-turut digabungkan sampai struktur cluster yang diinginkan diperoleh.
   - Divisive hierarchical clustering -Semua objek awalnya milik satu cluster. Kemudian cluster dibagi menjadi sub-cluster, yang secara berturut-turut dibagi menjadi sub-cluster mereka sendiri. Proses ini berlanjut hingga struktur kluster yang diinginkan diperoleh.

   Hasilnya merupakan metode hirarkis bersarang pada kelompok objek dan tingkat kesamaan di mana perubahan clustering. Clustering objek data diperoleh dengan memotong dendrogram pada tingkat kesamaan yang diinginkan. Penggabungan atau pembagian cluster dilakukan menurut beberapa ukuran kesamaan, dipilih untuk mengoptimalkan beberapa kriteria. Metode clustering hirarkis dapat dibagi lagi sesuai dengan cara pengukuran kesamaan dihitung [11].
b. Clustering Partisi: Metode partisi memindahkan instance dengan memindahkannya dari satu cluster ke cluster lainnya, dimulai dari partisi awal. Metode semacam itu biasanya mengharuskan jumlah klaster akan ditetapkan sebelumnya oleh pengguna. Untuk mencapai optimalisasi global dalam pengelompokan berbasis partisi, proses enumerasi yang lengkap dari semua partisi yang mungkin diperlukan. Karena ini tidak layak, heuristik tamak tertentu digunakan dalam bentuk pengoptimalan berulang. Yakni, metode relokasi memindahkan secara berulang-ulang poin antara cluster.

c. Clustering berdasarkan kompleksitas : metode berbasis Kepadatan berasumsi bahwa titik-titik yang termasuk dalam setiap cluster diambil dari distribusi probabilitas tertentu [12]. Distribusi data secara keseluruhan diasumsikan sebagai campuran dari beberapa distribusi. Tujuan dari metode ini adalah untuk mengidentifikasi cluster dan parameter distribusinya. Metode ini dirancang untuk menemukan kelompok bentuk acak yang tidak selalu cembung. Idenya adalah untuk mengelompokkan pertumbuhan yang diberikan untuk kepadatan (jumlah objek atau titik data) di lingkungan yang melebihi beberapa ambang batas. Yakni, lingkungan radius tertentu harus mengandung setidaknya sejumlah objek minimum. Ketika setiap cluster dicirikan oleh mode lokal atau maksimal dari fungsi kepadatan, metode ini disebut mode pencarian.
d. Clustering berbasis jaringan: Metode ini membagi ruang menjadi jumlah sel yang terbatas yang membentuk struktur grid di mana semua operasi untuk clustering dilakukan. Keuntungan utama dari pendekatan ini adalah waktu pemrosesan yang cepat [2].

2.1. Agglomerative Hierarchical Clustering (AHC)

Agglomerative Hierarchical Clustering (AHC) adalah metode clustering bersifat bottom-up yaitu menggabungkan cluster-cluster menjadi satu cluster tunggal. Agglomerative hierarchical clustering (AHC) dengan menggunakan buttom –up , dimulai dari masing
-masing data sebagai sebuah cluster, kemudian secara rekrusif mencar kelompok terdekat sebagai pasangan yang kemudian akan digabungkan menjadi kelompok yang lebih besar. Proses tersebut diulang terus sehingga tampak bergerak keatas membentuk hirarki [13]. Langkah Algoritma Agglomerative Hierarchical Clustering :

a. Hitung Matrik Jarak antar data.
b. Gabungkan dua kelompok terdekat berdasarkan parameter kedekatan yang ditentukan.
c. Perbarui Matrik Jarak antar data untuk merepresentasikan kedekatan diantara kelompok baru dan kelompok yang masih tersisa.
d. Ulangi langkah b dan c higga hanya satu kelompok yang tersisa.

Salah satu algoritma yang digunakan dalam agglomerative *hierarchical clustering* adalah *complete linkage. Complete linkage* terjadi apabila kelompok- kelompok digabung berdasarkan jarak terjauh dari satu cluster dengan *cluster* lainnya. Persamaan *Agglomerative Hierarchical Clustering (*AHC) dengan menggunakan Manhattan Distance :

Persamaan umum :

$$D_{man}(x,y) = \sum_{j=1}^{d} |x_j - y_j| \quad (1)$$

Euclidian Distance :

$$D(x_2, x_1) = \sqrt{\sum_{j=1}^{d} |x_{2j} - y_{1j}|2} \quad (2)$$

Beberapa metode pengelompokkan secara Hierarki Aglomeratif :

Single Linkage (Jarak Terdekat)

$$d_{uv=min\{d_{uv}\},d_{uv}\,\epsilon D} \quad (3)$$

Complete Linkage (Jarak Terjauh)

$$d_{uv=max\{d_{uv}\},d_{uv}\,\epsilon D} \quad (4)$$

Average Linkage (Jarak Rata-rata)

$$d_{uv=average\{d_{uv}\},d_{uv}\,\epsilon D} \quad (5)$$

Pada perhitungan data yang akan dihitung menggunakan *Agglomerative Hierarchical Clustering* (AHC)- menggunakan metode *Complete Linkage*, kedekatan diantara empat data cluster ditentukan oleh jarak terjauh dari empat cluster yang berbeda. Dengan cara ini maka akan dimulai dari masing-masing data sebagai cluster, kemudian mencari tetangga terjauh dengan menggunakan *complete linkage* untuk menggabungkan empat cluster berikutnya hingga semuanya menjadi dua cluster.

2.2. Data Penelitian

Tahapan penelitian untuk penerpan metode yang digunakan dalam menunjang pembuatan program *clustering* tingkat kesehatan masyarakat dengan menggunakan metode *Agglomerative Hierarchical Clustering- Complete Linkage* (AHC). Data dipilih dari data jarak terjauh yang diambil dari nilai empat data indikator derajat kesehatan masyarakat. Dalam menentukan perhitungan tingkat kesehatan masyarakat menggunakan metode *Complete Linkage* (Jarak Terjauh). Contoh data diambil dari profil kesehatan dari lima wilayah di suatu daerah tertentu. Data nilai empat indikator derajat kesehatan masyarakat dituliskan pada Tabel 1.

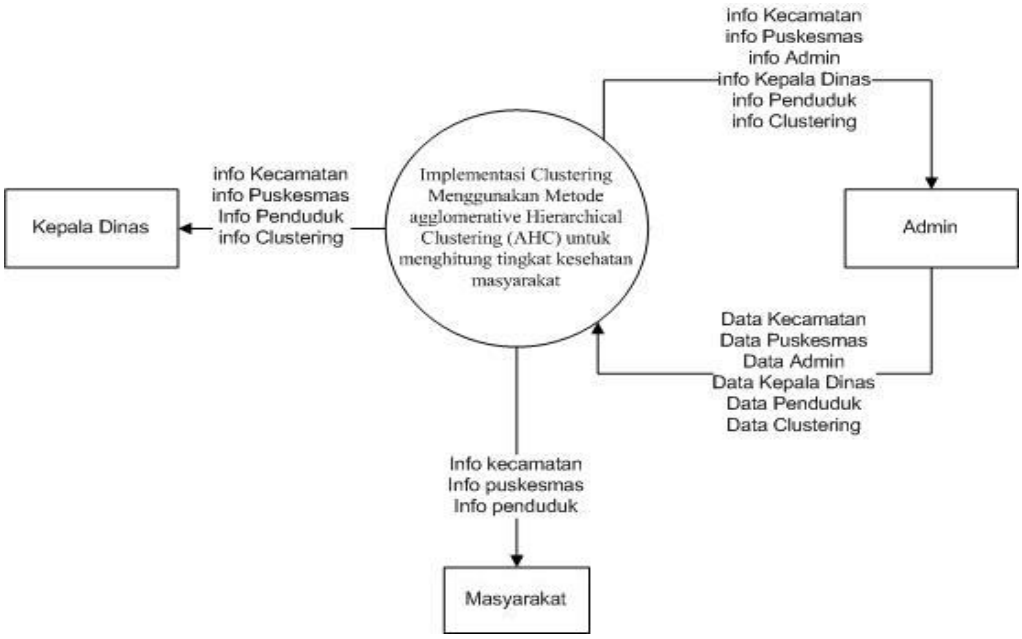
Tabel 1. Sampel Data Perhitungan

D_{man}	1 Kematian Bayi	2 Kematian Balita	3 Kematian Ibu	4 Kelahiran Kasar
1	2	2	3	32,451
2	1	2	1	29,426
3	2	0	0	35,398
4	2	2	0	29,405
5	2	0	1	16,423

3. Pembahasan

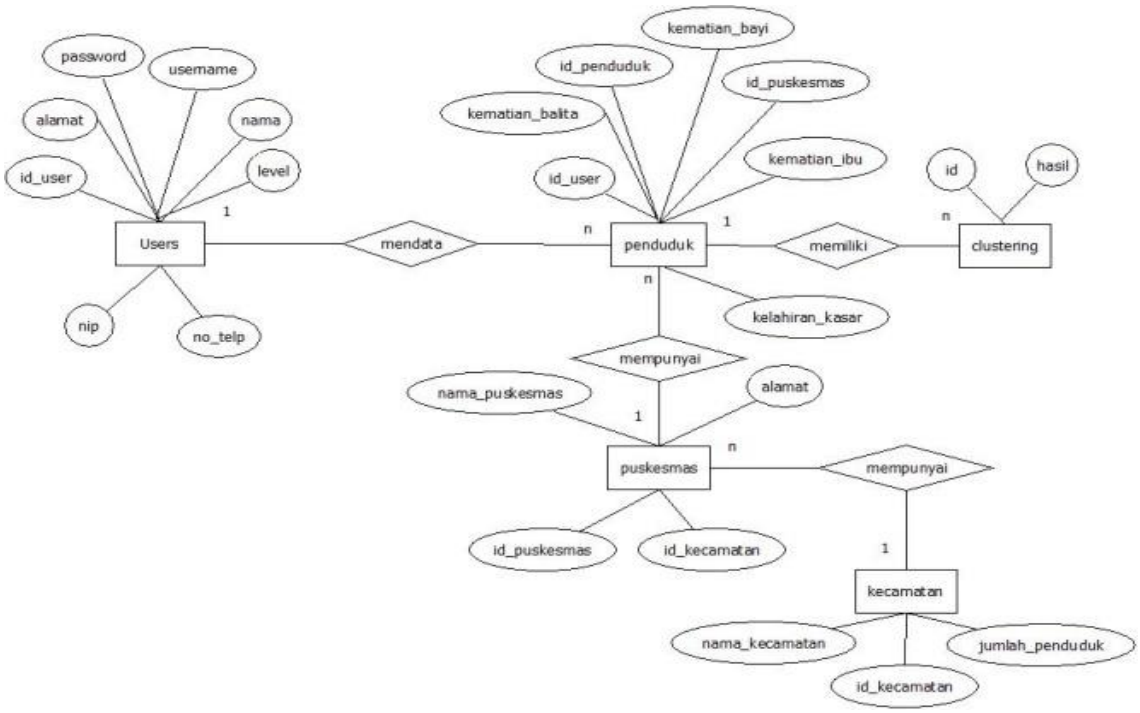
3.1. Analisis dan Perancangan Sistem

Analisis sistem yang dilakukan oleh penulis mencakup analisis proses sistem dan analisis kebutuhan sistem. Ini dilakukan untuk mengetahui kebutuhan input data, kebutuhan proses, dan keluaran sistem. Gambar 1 adalah diagram konteks merupakan bentuk Diagram Alir Data (DAD) yang paling sederhana. Admin adalah pengguna penuh sistem yang mampu mengakses sistem secara keseluruhan. Kepala dinas adalah relawan yang membatu dan memiliki hak akses terhadap sistem namun dibatasi sehingga ada menu yang hanya dapat di akses dan digunakan oleh kepala dinas.



Gambar 1. Diagram Konteks

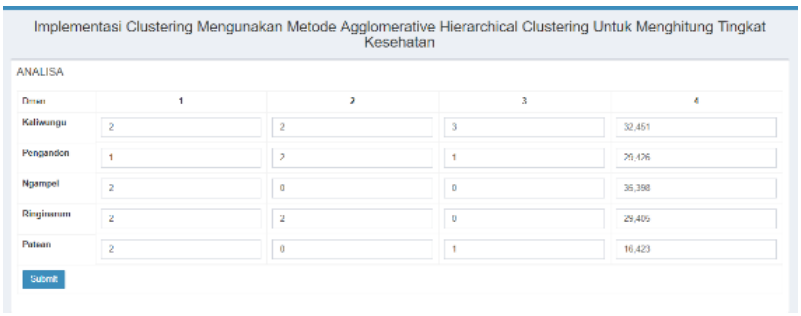
Dalam perancangan *Entity Relationship Diagram* (ERD) pada sistem ini menggunakan lima entitas, yaitu Users, Penduduk, Clustering, Puskesmas, dan Kecamatan. Gambar 2. Menunjukkan entitas-entitas yang digunakan dalam perancangan ERD pada sistem ini akan saling berelasi antar entitas. Relasi-relasi yang terjadi antar entitas.



Gambar 2. *Entity Relationship Diagram* (ERD)

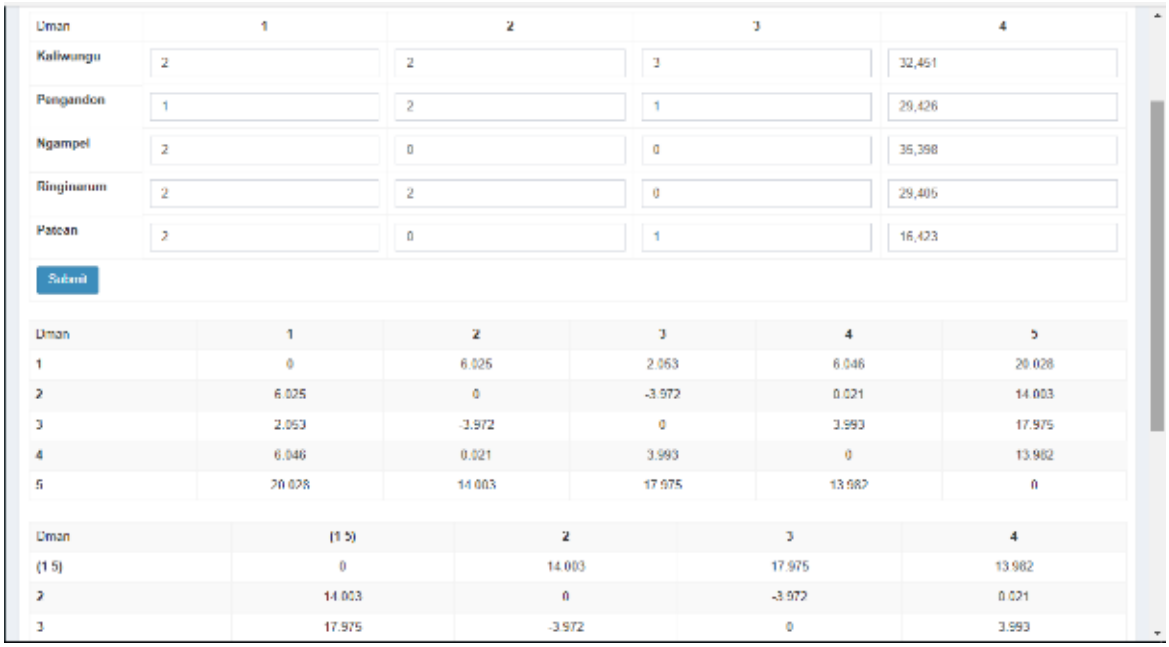
3.2. Analisis Clustering

Menu Analisis Cluster merupakan menu untuk menampilkan data tingkat kesehatan masyarakat yang dari empat nilai indikator derajat kesehatan yang didapatkan dari data kecamatan dan data puskesmas ditunjukkan pada Gambar 3.



Gambar 3. Menu Analisis Clustering

Hasil perhitungan clustering menggunakan metode *Agglomerative Hierarchical Clustering* (AHC). Data yang ada pada Gambar 4 menunjukkan hasil perhitungan bahwa nilai yang diambil dari nilai empat indikator derajat kesehatan yang dihasilkan dari nilai kematian bayi, kematian balita, kematian ibu dan kelahiran kasar.



Uman	1	2	3	4
Kaliwungu	2	2	3	32,461
Pengandon	1	2	1	29,426
Ngampel	2	0	0	35,398
Ringinarum	2	2	0	29,405
Patean	2	0	1	16,423

Submit

Uman	1	2	3	4	5
1	0	6.025	2.053	6.046	20.028
2	6.025	0	-3.972	0.021	14.003
3	2.053	-3.972	0	3.993	17.975
4	6.046	0.021	3.993	0	13.982
5	20.028	14.003	17.975	13.982	0

Uman	(1 5)	2	3	4
(1 5)	0	14.003	17.975	13.982
2	14.003	0	-3.972	0.021
3	17.975	-3.972	0	3.993

Gambar 4. Hasil Perhitungan Agglomerative Hierarchical Clustering

4. Kesimpulan

Dari hasil pengujian dan analisis sistem klasterisasi data kesehatan penduduk untuk menentukan tingkat kesehatan masyarakat dengan menggunakan metode *Agglomerative Hierarchical Clustering* (AHC), maka dapat disimpulkan hal-hal sebagai berikut:

a. Hasilnya, menggabungkan dari semua klaster menjadi dua klaster besar untuk melihat hasil yang sehat dan tidak sehat.
b. Metode *Agglomerative Hierarchical Clustering* (AHC)-*Complete Linkage* (Jarak terjauh) dalam sistem yang telah dibuat dapat mengatasi kerumitan dalam menghitung jarak terjauh, namun belum bisa menampilkan hasil pohon atau grafiknya yang menentukan sehat atau tidak sehat dalam hasil akhirnya.

Daftar Pustaka

[1] Mesakar, S. S., & Chaudhari, M. S. (2012). Review Paper On Data Clustering Of Categorical Data. *International Journal of Engineering Research & Technology (IJERT)*, *1*(10), 1-18.

[2] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[3] Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge & Data Engineering*, (11), 1370-1386.

[4] Giannotti, F., Gozzi, C., & Manco, G. (2002, August). Clustering transactional data. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 175-187). Springer, Berlin, Heidelberg.

[5] Mathieu, R. G., & Gibson, J. E. (1993). A methodology for large-scale R&D planning based on cluster analysis. *IEEE Transactions on Engineering Management*, *40*(3), 283-292.

[6] Haimov, S., Michalev, M., Savchenko, A., & Yordanov, O. I. (1989). Classification of radar signatures by autoregressive model fitting and cluster analysis. *IEEE Transactions on Geoscience and Remote Sensing*, *27*(5), 606-610.
[7] Yanto, I. T. R., Herawan, T., & Deris, M. M. (2011). Data clustering using variable precision rough set. *Intelligent Data Analysis*, *15*(4), 465-482.
[8] Yanto, I. T. R., Vitasari, P., Herawan, T., & Deris, M. M. (2012). Applying variable precision rough set model for clustering student suffering study's anxiety. *Expert Systems with Applications*, *39*(1), 452-459.
[9] Herawan, T. (2012). Rough clustering for cancer datasets. In *International Journal of Modern Physics: Conference Series* (Vol. 9, pp. 240-258). World Scientific Publishing Company.
[10] Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, *41*(8), 578-588.
[11] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.
[12] Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803-821.
[13] Prasetyo, E. (2012). Data Mining konsep dan Aplikasi menggunakan MATLAB. *Yogyakarta: Andi*.