



Studi Kasus

PENERAPAN *DATA MINING* DALAM PENGELOMPOKAN PENDERITA *THALASSAEMIA*

Heni Sulastri¹, Acep Irham Gufroni¹

¹Jurusan Informatika, Universitas Siliwangi, Tasikmalaya 46115, Indonesia

INFORMASI ARTIKEL

Sejarah Artikel:

Diterima Redaksi: 12 Juli 2017

Revisi Akhir: 01 September 2017

Diterbitkan Online: 26 September 2017

KATA KUNCI

Analysis

Clustering

Data Mining

K-Means

Thalassaemia

KORESPONDENSI

Telepon: +62(265) 330634

E-mail: henisulastri@unsil.ac.id

A B S T R A C T

Thalassaemia is the genetic disease caused by deficiency and synthesis of globin chains. It influences our body by decreasing erythrocyte and hemoglobin degree. People with Thalassaemia in 2015 at Tasikmalaya, Garut, and Ciamis West Java were 203 people. They organized in POPTI Tasikmalaya branch that placed in Dr. Soekardjo and Prasetya Bunda hospital. On the therapy process, they have different time needs and blood volume needs in every transfusion process. On the other hand, the difference transfusion levels also influence in giving iron chelation medicine. Furthermore, the method needed to help POPTI committee and health staff in appropriating blood volume and Iron Chelating Agent trough Thalassaemia people. Data mining method used by applying clustering method used K-means algorithm. Furthermore, this research conducted to categorized people with Thalassaemia based on blood volume need and HB in every transfusion process. Moreover, the pattern known by minor Thalassaemia, intermediate Thalassaemia, and major Thalassaemia based on age pattern, HB level in transfusion process, and blood volume needs. The research method is begin by pre observation and data mining analysis method to analyze data on data mining using 3 steps of KDD such as data cleaning, data integration, data selection, data transformation, and data knowledge presentation. Further, the result of this research has 374 data that divided into 3 cluster. They are cluster 1 that has 214 data, cluster 2 has 137 data, and cluster 3 that has 23 data with the pattern that shows that the transfusion blood volume increase based on patient's age.

1. PENDAHULUAN

Data Mining 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. Data Mining dapat diterapkan pada berbagai bidang yang mempunyai sejumlah data. Menurut Daryl Pregibon bahwa "Data mining adalah campuran dari statistic, kecerdasan buatan, dan riset basis data" yang masih berkembang [1].

Dalam dunia kesehatan data mining dapat dimanfaatkan untuk mendapatkan informasi seperti penentuan kriteria suatu penyakit, misalnya tingkat kebutuhan transfusi darah dari penderita Thalassaemia. Kebutuhan waktu transfusi yang berbeda-beda dari setiap penderita menjadi sebuah masalah dalam mempersiapkan pemberian jumlah obat terafi kelas besi dan kesiapan pendonor darah.

Di wilayah Kota dan Kabupaten Tasikmalaya penderita Thalassaemia berjumlah 203 penderita pada tahun 2015. Dan dari setiap penderita Thalassaemia tersebut memiliki jadwal transfusi darah yang berbeda-beda, yaitu 2 minggu sekali, 3 minggu sekali, 1 bulan sekali atau lebih, hasil tersebut berdasarkan data yang diperoleh dari Perhimpunan Orang Tua Penderita Thalassaemia Indonesia (POPTI) berdasarkan catatan data rekam medik di RSUD dr. Soekardjo dan RS. Prasetya Bunda.

Berdasarkan uraian diatas, untuk mengetahui klusterisasi atau pengelompokan waktu dan kebutuhan transfusi darah dari penderita Thalassaemia yang ada di wilayah Tasikmalaya, maka akan dilakukan penelitian mengenai *Clustering Analysis* menggunakan Algoritma *K-Means* dengan Pendekatan Data Mining Untuk pengelompokan penderita Thalassaemia di POPTI Cabang Tasikmalaya. Dengan harapan dapat memberikan pengetahuan berupa tabel dan grafik mengenai kelompok atau kriteria dari penderita Thalassaemia di Tasikmalaya sehingga dapat menjadi acuan bagi dokter, perawat dan orang tua dalam

memberikan terapi obat dan kebutuhan transfusi darah dari setiap penderita.

Batasan masalah pada penelitian ini yaitu : tempat studi kasus Perhimpunan Orang Tua Penderita Thalassaemia Indonesia (POPTI) Cabang Tasikmalaya dengan menggunakan data rekam medik pasien Thalassaemia di wilayah Kota dan Kabupaten Tasikmalaya yang dirawat di RSUD dr Soekardjo dan RS Prasetya Bunda pada tahun 2015. Analisis dilakukan dengan menggunakan Algoritma *K-Means Clustering*. Dan analisis digunakan untuk mencari informasi dari data rekam medik pasien Thalassaemia di wilayah kota dan kabupaten Tasikmalaya yang ada sebagai penentuan tingkat kebutuhan transfusi darah penderita Thalassaemia.

Tujuan pada penelitian ini yaitu untuk mengelompokkan data kriteria penderita Thalassaemia berdasarkan umur, Hb level dan kebutuhan jumlah darah dengan pendekatan data mining menggunakan algoritma *K-means*. Sehingga memberikan informasi dan pengetahuan berupa tabel dan grafik tentang pola kriteria dari penderita penyakit Thalassaemia.

Manfaat pada penelitian ini yaitu tersedianya informasi mengenai pengelompokan kriteria penderita Thalassaemia di Kota dan Kabupaten Tasikmalaya dan membantu pihak-pihak yang berkepentingan untuk mengetahui kriteria penderita Thalassaemia yang ada di Kota Tasikmalaya. Dan menjadi rujukan tenaga medis dalam penanganan pemberian terapi bagi penderita Thalassaemia.

2. TINJAUAN PUSTAKA

2.1. Thalassaemia

Thalasemia merupakan penyakit genetik yang ditandai dengan keadaan yang disebabkan oleh defisiensi atau tidak adanya sintesis rantai globin. Thalassaemia dapat mempengaruhi tubuh dengan menurunkan kadar sel darah merah dan hemoglobin dari angka normal. Hemoglobin merupakan protein kaya berperan dalam distribusi oksigen maupun karbondioksida dalam tubuh. Orang yang menderita thalasemia dapat mengalami anemia berat ataupun ringan. Hal ini dikarenakan rendahnya jumlah sel darah merah atau tidak cukupnya hemoglobin pada sel darah merah [10].

Menurut Potts dan Mendleco tahun 2007 Thalassaemia merupakan kelainan seumur hidup yang diklasifikasikan menjadi Thalassaemia alfa dan Thalassaemia beta berdasarkan sintesis rantai globin dalam hemoglobin yang mengalami gangguan, dan jenis Thalassaemia mayor dan Thalassaemia minor tergantung jumlah gen yang mengalami kerusakan/kecacatan [3].

Jenis Thalassaemia berdasarkan jumlah gen yang mengalami gangguan diklasifikasikan menjadi Thalassaemia minor, Thalassaemia intermedia dan Thalassaemia mayor.

Pengobatan untuk penyakit Thalassaemia tergantung pada tipe thalasemia dan berat tidaknya Thalassaemia yang diderita. Penderita dengan Thalassaemia yang ringan atau bersifat asimtomatik biasanya hanya mendapat sedikit

perawatan/pengobatan bahkan tidak mendapat pengobatan sama sekali. Ada tiga pengobatan yang paling sering digunakan oleh dokter. Ketiganya itu antara lain transfusi darah, terapi iron chelation, dan suplemen asam folat.

2.2. Data Mining

Menurut [2] mengatakan bahwa “Data mining merupakan disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data. Menurut [9] data mining adalah suatu metode pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan metode data mining ini dapat digunakan untuk mengambil keputusan di masa depan. Data mining ini juga dikenal dengan istilah *pattern recognition* [8].

[2] Menyebutkan bahwa KDD atau *Knowledge Discovery from Data*, merupakan proses terstruktur, yaitu sebagai berikut:

1. **Data Cleaning** adalah Proses membersihkan data dari data noise dan tidak konsisten.
2. **Data Integration** adalah Proses untuk menggabungkan data dari beberapa sumber yang berbeda.
3. **Data Selection** adalah Proses untuk memilih data dari database yang sesuai dengan tujuan analisis.
4. **Data Transformation** adalah Proses mengubah bentuk data menjadi data yang sesuai untuk proses Mining.
5. **Data Mining** adalah Proses penting yang menggunakan sebuah metode tertentu untuk memperoleh sebuah pola dari data.
6. **Pattern Evaluation** adalah Proses mengidentifikasi pola.
7. **Knowledge Presentation** adalah yang dapat merepresentasikan informasi yang dibutuhkan, proses dimana informasi yang telah didapatkan kemudian digunakan oleh pemilik data.

2.3. Clustering

Menurut [8] *clustering* merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity*) antara satu data dengan data yang lain. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data atau obyek ke dalam *cluster (group)* sehingga dalam setiap *cluster* dapat berisi data yang semirip mungkin. Dalam *clustering* metode ini berusaha untuk menempatkan obyek yang mirip (jaraknya dekat) dalam satu *cluster* dan membuat jarak antar *cluster* sejauh mungkin. Ini berarti obyek dalam satu *cluster* sangat mirip satu dengan lain dan berbeda dengan obyek dalam *cluster-cluster* yang lain [8].

Dalam *data mining* ada dua jenis metode *clustering* yang digunakan dalam pengelompokan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering*. Metode *non-hierarchical (partitioning) clustering* dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang diinginkan (dua *cluster*, tiga *cluster*, atau lain sebagainya). Setelah jumlah *cluster* diketahui, baru proses *cluster* dilakukan. Metode ini biasa disebut dengan *K-Means Clustering* [9].

2.4. Algoritma K-Means

Menurut [13], Algoritma *K-Means* merupakan Algoritma pengelompokan iteratif yang melakukan partisi set data kedalam sejumlah *K cluster* yang sudah ditetapkan diawal. Algoritma *K-Means* sederhana untuk diimplementasikan dan dijalankan, relative cepat, mudah beradaptasi, umum penggunaannya dalam praktek. Secara historis, K-Means menjadi salah satu algoritma yang paling penting dalam bidang data mining [5].

Algoritma *K-means* adalah algoritma yang terbaik dalam algoritma *partitional clustering* dan yang paling sering digunakan diantara algoritma *clustering* lainnya karena kesederhanaan dan efisiensinya. Kelebihan Algoritma *K-means* itu sendiri menurut K. Arai and A. R. Barakbah, merupakan algoritma *clustering* yang paling sederhana dan umum, hal ini dikarenakan *K-means* mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien. Namun, *K-means* mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster* [11].

Algoritma *K-Means* dimulai dengan pemilihan secara acak *K*, *K* disini merupakan banyaknya *cluster* yang ingin dibentuk. Kemudian tetapkan nilai-nilai *K* secara random, untuk sementara nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan *centroid* secara random, *mean* atau "*means*". Hitung jarak setiap data yang ada terhadap masing-masing *centroid* menggunakan rumus *Euclidean* hingga ditemukan jarak yang paling dekat dari setiap data dengan *centroid*. Klasifikasikan setiap data berdasarkan kedekatannya dengan *centroid*. Lakukan langkah tersebut hingga nilai *centroid* tidak berubah [12].

Jarak *Euclidean* yang dirumuskan sebagai berikut:

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

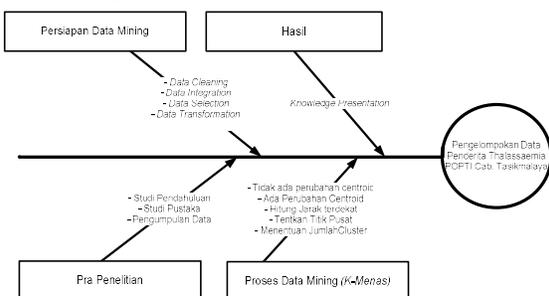
$D(i, j)$ = Jarak data ke *i* ke pusat *cluster j*

X_{ki} = Data ke *i* pada atribut dat ke *k*

X_{kj} = Data ke *j* pada atribut dat ke *k*

3. METODOLOGI

Tahapan-tahapan yang dilakukan dalam penelitian adalah seperti pada gambar dibawah ini:



Gambar 1. Diagram Penelitian

Penelitian ini dibagi menjadi tiga tahapan utama yaitu:

1. Pra penelitian, dimulai dengan melakukan studi pendahuluan untuk mengidentifikasi masalah yang ada, studi pustaka, dan pengumpulan data penderita

Thalassaemia di kabupaten dan Kota Tasikmalaya tahun 2015.

2. Persiapan proses data mining terdiri dari, *Data Cleaning*, *Data Autentication*, *Data Integration*, *Data Selection*, dan *Data Transformation*.
3. Proses Data Mining dengan mengelompokkan menggunakan algoritma K-means.

3.1. Pra Penelitian

1. Studi Pendahuluan ini dimulai dengan melakukan pengamatan secara langsung pada data penderita Thalassaemia. Pengamatan secara langsung ini dilakukan melalui wawancara terhadap pihak internal pengurus Perhimpunan Orang Tua Penderita Thalassaemia Indonesia (POPTI) Cabang Tasikmalaya.
2. Setelah rumusan masalah didapatkan, dilakukan studi pustaka mencari referensi dari jurnal, buku, dan *e-book* untuk mendapatkan teori-teori tentang permasalahan yang telah didapatkan, sebagai panduan dalam mendapatkan solusi untuk permasalahan yang telah ditemukan.
3. Untuk memperoleh data penderita Thalassaemia yang akan digunakan pada proses data mining, dilakukan pendekatan kepada pengurus POPTI guna mendapatkan data penderita Thalassaemia yang ada di kota dan kabupaten Tasikmalaya tahun 2015 yang sebenarnya.

3.2. Persiapan Data Mining

Preprocessing Data Mining dapat meningkatkan kualitas data, sehingga data yang diperoleh langsung dari pengurus POPTI cab. Tasikmalaya diolah terlebih dahulu melalui tahap-tahap *data cleaning*, *data integration*, *data selection*, dan *data transformation*. Hal tersebut dilakukan agar data yang diolah lebih berkualitas artinya data-data tersebut bersifat objektif, representatif, memiliki *sampling error* yang kecil, terbaharui dan relevan. Persiapan tersebut antara lain:

1. Data Cleaning

Data Cleaning merupakan proses untuk dapat mengatasi nilai yang hilang, *noise* dan data yang tidak konsisten [2].

2. Data Integration

Data Integration merupakan proses menggabungkan data dari banyak database. Setelah dilakukan data authentication terdapat data terpisah yaitu data tanggal lahir sehingga didapatkan umur penderita, maka dilakukan proses *cleaning* kedua dengan mengintegrasikan data awal penderita Thalassaemia.

3. Data Selection

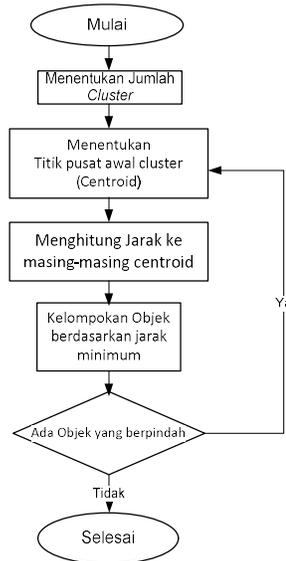
Data Selection merupakan proses meminimalkan jumlah data yang digunakan untuk proses *mining* dengan tetap merepresentasikan data aslinya [2]. *Data selection* dapat berupa *sampling*, *denoising*, dan *feature extraction*.

4. Data Transformation

Data transformation dilakukan untuk mengubah bentuk dan format data. Hal ini tentunya sangat membantu memudahkan pengguna dalam proses *mining* ataupun memahami hasil yang didapat [2].

Dalam proses *data transformation* bisa dilakukan dengan *centering*, *normalization*, dan *scaling*.

3.3. Proses Data Mining K-means



Gambar 2. Flowchart Algoritma K-means

Gambar 2 diatas menunjukkan langkah-langkah menentukan *clustering* dengan metode K-Means menurut [8] yang terdiri dari beberapa tahapan sebagai berikut:

1. Menentukan jumlah *cluster k*.
2. Melakukan alokasi data kedalam *cluster* secara random.
3. Melakukan alokasi semua data/objek ke *cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke *cluster* tertentu ditentukan jarak antara 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 jarak *Euclidean* yang dirumuskan sebagai berikut:

$$d_{ij} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \dots + (x_{ki} - x_{kj})^2} \dots (1)$$

Dimana:

d_{ij} = Jarak data ke i ke pusat *cluster j*

X_{ki} = Data ke- i pada atribut data ke- k

X_{kj} = Data ke- j pada atribut data ke- k

4. Melakukan perhitungan kembali pusat *cluster* 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.

Tugaskan lagi setiap objek 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.

3.4. Knowledge Presentation

Setelah dilakukan pengelompokan menggunakan algoritma K-means maka diperoleh *Knowledge Presentation* adalah proses yang dapat merepresentasikan informasi yang dibutuhkan, dimana informasi yang telah didapatkan kemudian digunakan.

4. HASIL DAN PEMBAHASAN

4.1. Data Cleaning

Data yang dianalisis yaitu data rekam medik penderita Thalassaemia di RS Dr. Soekardjo dan RS Prasetya Bunda panda tahun 2015. Data set yang didapatkan adalah sebanyak 2068 data pelaksanaan transfusi. Dari data awal kolom no, kolom golongan darah, kolom kadar feritin, kolom resep dokter, kolom realisasi obat, kolom BB, dan kolom keterangan dihilangkan karena tidak sesuai dengan kebutuhan proses perhitungan data mining. Sehingga kolom yang digunakan adalah kolom no, kolom tanggal transfusi, kolom nama, kolom Hb level, kolom volume darah.

Berdasarkan data tabel tersebut yang menjadi atribut utama adalah kolom Hb level dan volume darah maka proses *data cleaning* untuk menentukan data *premature* hanya terjadi pada kolom tersebut, dengan hasil :

Tabel 1. Data *Premature*

Kolom Hb Level	72 data
Kolom Volume Darah	4 data
Kolom Hb Level + Kolom Volume Darah	2 data

Totalnya adalah 78 *data premature* atau sekitar 3.77%. Sehingga data baru yang dihasilkan adalah 1990 data pelaksanaan transfusi.

4.2. Data Integration

Setelah dilakukan *data authentication* terdapat data terpisah yaitu data tanggal lahir sehingga didapatkan umur penderita, maka dilakukan proses *cleaning* kedua dengan mengintegrasikan data awal penderita Thalassaemia. Sementara untuk data *premature* pada data primer tidak ada data tambahan yang dapat melengkapi data *premature* tersebut.

4.3. Data Selection

Untuk proses *data selection* yang digunakan yaitu proses *sampling*, atau proses pengambilan contoh data. Dari 1990 data diambil *data sampling* berupa data setiap penderita yaitu data minimum dan data maksimum hb level setiap kali pelaksanaan transfusi. Sehingga diperoleh 374 data yang akan diproses

4.4. Proses Pengelompokan Algoritma K-Means

4.4.1. Menentukan Jumlah Cluster

Pada pengelompokan penderita Thalassaemia ini dibuat menjadi 3 *cluster*. Penentuan *cluster* tersebut terdiri dari 2 variabel yaitu *cluster* untuk variabel Hb level dan

cluster untuk variabel volume darah. Dengan cluster sebagai berikut:

- a. Cluster pertama yaitu kelompok Thalassaemia ringan,
- b. Cluster kedua yaitu kelompok Thalassaemia sedang, dan
- c. Cluster ketiga adalah kelompok Thalassaemia berat.

Penentuan kelompok tersebut berdasarkan informasi dari pengurus POPTI Cabang Tasikmalaya, yang menyatakan bahwa setiap penderita memiliki kebutuhan terhadap transfusi darah yang berbeda-beda berdasarkan Hb level dan volume darah yang dibutuhkan. Dan berdasarkan penelitian yang telah dilakukan [6].

4.4.2. Menentukan Titik Pusat Awal Cluster (Centroid)

Pada penelitian ini penentuan titik pusat awal berdasarkan proses perhitungan data yang paling banyak muncul (modus) dari kolom volume darah, kemudian dari tiga data yang muncul dihitung nilai rata-rata pada kolom Hb level. Proses penentuan titik pusat ini tidak dilakukan secara random karena pembangkitan awal titik pusat secara random itu K-means tidak menjamin hasil pengklasteran yang unik. Inilah yang menyebabkan metode K-means sulit untuk mencapai global optimum, akan tetapi hanya local minimum, maka besar kemungkinan ini akan menyebabkan hasil pengklasteran yang tidak tepat.

Tabel 2. Titik pusat awal (centroid)

Cluster	1	2	3
Volume Darah	200	400	600
Hb level	8,616	7,136	5,500

Tabel 2 merupakan titik pusat awal pada cluster. Penentuan cluster tersebut terdiri dari 2 variabel yaitu cluster untuk variabel Hb level dan cluster untuk variabel volume darah. Dengan cluster sebagai berikut:

1. Cluster pertama yaitu kelompok Thalassaemia ringan,
2. Cluster kedua yaitu kelompok Thalassaemia sedang, dan
3. Cluster ketiga adalah kelompok Thalassaemia berat. Penentuan kelompok tersebut berdasarkan informasi dari pengurus POPTI Cabang Tasikmalaya, yang menyatakan bahwa setiap penderita memiliki kebutuhan akan transfusi darah yang berbeda-beda berdasarkan Hb level dan volume darah yang dibutuhkan.

4.4.3. Menghitung Jarak ke Masing-masing Centroid dan Mengelompokkan Berdasarkan Jarak Minimum

Setelah didapat titik pusat awal cluster, kemudian dilakukan perhitungan jarak Euclidian, dan mengelompokkan berdasarkan jarak terkecil selanjutnya akan di dapat nilai centroid baru untuk acuan perhitungan berikutnya sampai nilai centroid sebelum dan sesudah bernilai sama.

Perhitungan jarak Euclidean pada iterasi 1

- 1) Berikut ini beberapa hasil perhitungan jarak Euclidean pada titik pusat cluster 1 (8,616; 200) :

$$D(1,1) = \sqrt{(7,7 - 8,616)^2 + (400 - 200)^2} = 200,002$$

$$D(1,2) = \sqrt{(5,5 - 8,616)^2 + (400 - 200)^2} = 200,024$$

$$D(1,3) = \sqrt{(9,6 - 8,616)^2 + (400 - 200)^2} = 200,002$$

<https://doi.org/10.25077/TEKNOSI.v3i2.2017.299-305>

$$D(1,4) = \sqrt{(5,7 - 8,616)^2 + (600 - 200)^2} = 400,011$$

$$D(1,5) = \sqrt{(10,9 - 8,616)^2 + (400 - 200)^2} = 200,013$$

Lakukan perhitungan yang sama sampai data ke 374.

- 2) Berikut ini beberapa hasil perhitungan jarak Euclidean pada titik pusat cluster 2 (7,136; 400):

$$D(2,1) = \sqrt{(7,7 - 7,136)^2 + (400 - 400)^2} = 0,564$$

$$D(2,2) = \sqrt{(5,5 - 7,136)^2 + (400 - 400)^2} = 1,636$$

$$D(2,3) = \sqrt{(9,6 - 7,136)^2 + (400 - 400)^2} = 2,464$$

$$D(2,4) = \sqrt{(5,7 - 7,136)^2 + (600 - 400)^2} = 200,005$$

$$D(2,5) = \sqrt{(10,9 - 7,136)^2 + (400 - 400)^2} = 3,764$$

Lakukan perhitungan yang sama sampai data ke 374.

- 3) Berikut ini beberapa hasil perhitungan jarak Euclidean pada titik pusat cluster 3 (5,500; 600):

$$D(3,1) = \sqrt{(7,7 - 5,500)^2 + (400 - 600)^2} = 200,012$$

$$D(3,2) = \sqrt{(5,5 - 5,500)^2 + (400 - 600)^2} = 200$$

$$D(3,3) = \sqrt{(9,6 - 5,500)^2 + (400 - 600)^2} = 200,042$$

$$D(3,4) = \sqrt{(5,7 - 5,500)^2 + (600 - 600)^2} = 0,200$$

$$D(3,5) = \sqrt{(10,9 - 5,500)^2 + (400 - 600)^2} = 200,073$$

Lakukan perhitungan yang sama sampai data ke 374.

Setelah di dapat nilai jarak Euclidean, kelompokkan data dengan pusat cluster yang memiliki jarak terkecil.

Tabel 3. Hasil Iterasi 1

Cluster	Titik Pusat Awal	Iterasi 1	
Cluster 1	8,616 200	8,619	197,430
Cluster 2	7,136 400	7,136	400
Cluster 3	5,500 600	5,504	617,391

Dari Tabel hasil iterasi 1 dapat dilihat centroid baru yang dihasilkan dari iterasi 1 dengan titik pusat awal tidak sama, maka perlu dilakukan iterasi ke dua.

Perhitungan jarak Euclidean pada iterasi 2

- 1) Berikut ini beberapa hasil perhitungan jarak Euclidean pada titik pusat cluster 1 (8,619; 197,430) :

$$D(1,1) = \sqrt{(7,7 - 8,619)^2 + (400 - 197,430)^2} = 202,572$$

$$D(1,2) = \sqrt{(5,5 - 8,619)^2 + (400 - 197,430)^2} = 202,594$$

$$D(1,3) = \sqrt{(9,6 - 8,619)^2 + (400 - 197,430)^2} = 202,572$$

$$D(1,4) = \sqrt{(5,7 - 8,619)^2 + (600 - 197,430)^2} = 402,581$$

$$D(1,5) = \sqrt{(10,9 - 8,619)^2 + (400 - 197,430)^2} = 202,583$$

Lakukan perhitungan yang sama sampai data ke 374.

- 2) Berikut ini beberapa hasil perhitungan jarak Euclidean pada titik pusat cluster 2 (7,136; 400) :

$$D(2,1) = \sqrt{(7,7 - 7,136)^2 + (400 - 400)^2} = 0,564$$

$$D(2,2) = \sqrt{(5,5 - 7,136)^2 + (400 - 400)^2} = 1,636$$

$$D(2,3) = \sqrt{(9,6 - 7,136)^2 + (400 - 400)^2} = 2,464$$

$$D(2,4) = \sqrt{(5,7 - 7,136)^2 + (600 - 400)^2} = 200,005$$

$$D(2,5) = \sqrt{(10,9 - 7,136)^2 + (400 - 400)^2} = 3,764$$

Lakukan perhitungan yang sama sampai data ke 374.

- 3) Berikut ini beberapa hasil perhitungan jarak Euclidean pada titik pusat cluster 3 (5,504; 617,391) :

$$D(3,1) = \sqrt{(7,7 - 5,504)^2 + (400 - 617,391)^2} = 217,402$$

$$D(3,2) = \sqrt{(5,5 - 5,504)^2 + (400 - 617,391)^2} = 217,391$$

$$D(3,3) = \sqrt{(9,6 - 5,504)^2 + (400 - 617,391)^2} = 217,430$$

$$D(3,4) = \sqrt{(5,7 - 5,504)^2 + (600 - 617,391)^2} = 17,392$$

$$D(3,5) = \sqrt{(10,9 - 5,504)^2 + (400 - 617,391)^2} = 217,458$$

Lakukan perhitungan yang sama sampai data ke 374.

Setelah di dapat nilai jarak Euclidean, kelompokkan data dengan pusat cluster yang memiliki jarak terkecil.

Tabel 4. Hasil Iterasi 2

Cluster	Iterasi 1	Iterasi 2	
Cluster 1	8,619 197,430	8,169	197,430
Cluster 2	7,136 400	7,136	400
Cluster 3	5,504 617,391	5,504	617,391

Dari Tabel 4.4 hasil iterasi 2 memiliki nilai centroid atau titik pusat awal yang sama dengan hasil dari iterasi 1. Maka proses perhitungan data mining dihentikan sampai iterasi 2.

Setelah dilakukan proses perhitungan ditemukan anggota-anggota dari setiap cluster yang terbentuk. Anggota-anggota tersebut dikelompokkan berdasarkan kemiripan data Hb level dan volume darah dengan *centroid* akhir dari setiap cluster. Maka diperoleh data sebagai berikut:

Tabel 5. Jumlah Data Hasil Pengelompokan Setiap Cluster

Cluster	Jumlah Data
Cluster 1	214
Cluster 2	137
Cluster 3	23
Total	374

4.4.4. Knowledge Presentation

1) Cluster 1

Berdasarkan hasil perhitungan algoritma *k-means* pada *cluster 1* ini diperoleh banyaknya data penderita Thalassaemia berjumlah 214 atau 57,219% dari total semua data yang berjumlah 374 data. Dari 214 data yang termasuk kedalam *cluster 1* diketahui bahwa volume darah (dalam cc) yang dibutuhkan adalah 100, 150 dan 200. Maka didapatkan pola Hb level, dan juga umur penderita sebagai berikut:

Tabel 6. Pola Umur dan Volume Darah Cluster 1

No	Kelompok Umur	Kelompok Hb Level	Frekuensi
1	0-5 Tahun	3,5 – 6,9	18
	Tahun	7,0 – 13,1	72
2	6–10 Tahun	5,1 – 6,8	10
	Tahun	7,0 – 12,1	46
3	11-15 Tahun	4,8 – 6,8	4
	Tahun	7,0 – 12,0	36
4	16-28 Tahun	5,8 – 6,9	1
	Tahun	7,0 – 10,5	10
5	Tidak Diketahui	4,9 – 6,9	6
		7,0 – 10,8	11
Jumlah			214

Dari tabel diatas diketahui bahwa data yang memiliki frekuensi paing banyak adalah data dengan Hb level antara 7,0 – 13,1 gr/dl dan umur antara 6 – 11 tahun.

2) Cluster 2

Berdasarkan hasil perhitungan algoritma *k-means* pada *cluster 2* ini diperoleh banyaknya data penderita Thalassaemia berjumlah 137 atau 36,63% dari total semua data yang berjumlah 374 data. Dari 137 data yang termasuk kedalam *cluster 2* diketahui bahwa volume darah (dalam cc) yang dibutuhkan adalah sekitar 400cc. Maka didapatkan pola Hb level, dan juga umur penderita sebagai berikut:

Tabel 7. Pola Umur dan Volume Darah Cluster 2

No	Kelompok Umur	Kelompok Hb Level	Frekuensi
1	0-5 Tahun	4,0 – 6,9	7
2	6–10 Tahun	4,0 – 6,9	19
		7,0 – 10,9	13
3	Tahun	4,0 – 6,9	24

4	11-15 Tahun	7,0 – 10,0	30
	16-28 Tahun	4,1 – 6,2	8
5	21-41 Tahun	7,0 – 12,1	14
		5,5 – 6,6	5
6	Tidak Diketahui	7,1 – 9,6	12
		4,6 – 6,1	3
Jumlah			137

Berdasarkan tabel diatas diketahui bahwa data yang memiliki frekuensi paing banyak adalah data dengan Hb level antara 7,0 – 10,0 gr/dl dan umur antara 11 – 15 tahun.

3) Cluster 3

Berdasarkan hasil perhitungan algoritma *k-means* pada *cluster 2* ini diperoleh banyaknya data penderita Thalassaemia berjumlah 23 atau 6,15% dari total semua data yang berjumlah 374 data. Dari 23 data yang termasuk kedalam *cluster 3* diketahui bahwa volume darah (dalam cc) yang dibutuhkan adalah 600 dan 800 cc. Maka didapatkan pola Hb level, dan juga umur penderita sebagai berikut :

Tabel 8. Pola Umur dan Volume Darah Cluster 3

No	Kelompok Umur	Kelompok Hb Level	Frekuensi
1	10-15 Tahun	3,8 – 6,0	7
2	16-20 Tahun	3,5 – 6,9	10
3	21-41 Tahun	5,0 – 8,2	4
6	Tidak Diketahui	6,0 – 6,2	2
Jumlah			23

Dari tabel diatas diketahui bahwa data yang memiliki frekuensi paling banyak adalah data dengan Hb level antara 3,5 – 6,9 gr/dl dengan umur antara 16 - 20 tahun.

Maka dengan adanya penelitian pengelompokan penderita Thalassaemia di POPTI Cabang Tasikmalaya ini dapat bermanfaat, yaitu sebagai berikut:

1. Orang Tua / Penderita

Menjadi bahan acuan untuk mempertahankan kadar Hb setiap kali transfusi. Setiap kali melakukan transfusi kadar Hb harus berada minimal diatas 50% kadar Hb normal atau lebih dari 7,0 gr/dl untuk dapat mempertahankan kualitas limpa penderita.

2. Pengurus POPTI Cab. Tasikmalaya

Menjadi bahan pertimbangan prioritas membantu orang tua/penderita dalam hal pengadaan darah untuk kebutuhan transfusi.

3. Tenaga Medis

Penderita Thalassaemia akan mengalami penumpukan zat besi akibat transfusi darah yang terus menerus. Maka diperlukan obat yang dapat membuang kadar zat besi dalam tubuh yaitu pemberian obat klasi besi berupa *desferal*, *perifox*, atau *exjade*. Sehingga dengan penelitian ini dapat menjadi acuan terhadap pemberian jumlah obat kepada penderita. Sehingga menjaga penderita Thalassaemia agar tidak segera dilakukan operasi pengangkatan limpa terutama pada penderita cluster 3 dengan jumlah transfusi 1 bulan 2 kali.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan penelitian ini didapat kesimpulan sebagai berikut:

1. Telah berhasil dilakukan pengelompokan data penderita Thalassaemia menggunakan algoritma K-means clustering menjadi 3 kelompok, yaitu untuk kelompok 1 berjumlah 214 data, kelompok 2 berjumlah 138 data, dan kelompok 3 berjumlah 23 data.
2. Dari setiap kelompok diperoleh pola data dengan frekuensi terbanyak yaitu :
 - a. Kelompok 1: data dengan Hb level antara 7,0 – 13,1 gr/dl dan umur antara 0 – 11 tahun dengan jumlah 72 data.
 - b. Kelompok 2: data dengan Hb level antara 7,0 – 10,0 gr/dl dan umur antara 11 – 15 tahun dengan jumlah 30 data.
 - c. Kelompok 3: data dengan Hb level antara 3,5 – 6,9 gr/dl dengan umur antara 16 - 20 tahun dengan jumlah 10 data.
3. Penentuan titik pusat awal (centroid) sangat berpengaruh terhadap jumlah iterasi yang akan dihitung.
4. Berdasarkan penelitian ini dapat dijadikan acuan yang dapat dilakukan baik oleh orang tua/penderita, pengurus POPTI, dan pihak tenaga medis. Mulai dari menjadi acuan pelaksanaan transfusi dan menjaga kadar Hb, persediaan labu darah, pemberian obat klasi besi, dan rekomendasi untuk operasi pengangkatan limpa.

5.2. Saran

1. Sebaiknya untuk penelitian selanjutnya dilakukan tindak lanjut untuk *data premature* yang diperoleh dari proses *data cleaning*.
2. Sebaiknya dilakukan analisis dan perhitungan yang sama dengan menggunakan Algoritma *clustering data mining* yang berbeda sehingga dapat dilakukan perbandingan terhadap hasil yang didapatkan.
3. Penelitian ini dapat dikembangkan menjadi penelitian dengan pembuatan aplikasi sehingga proses perhitungan data mining dapat dilakukan secara *up to date*.

DAFTAR PUSTAKA

- [1] Gorunescu, F. (2011). *Data Mining : Concepts, Models and Techniques*. New York: Springer-Verlag.
- [2] Han, J. Kamber, M & Jian, Pei. *Data Mining : Concepts and techniques, Third Edition*. America: Morgan Kauffman, San Francisco, 2011.
- [3] Indriati, Ganis. (2011). *Pengalaman Ibu dalam Merawat Anak Dengan Thalassaemia di Jakarta*. Tesis Program Magister Fakultas Keperawatan Universitas Indonesia. [Online]. Tersedia: <http://lib.ui.ac.id/file?file=digital/20280932-T%20Ganis%20Indriati.pdf> [3 Mei 2016]
- [4] Peraturan Menteri Kesehatan Republik Indonesia. (2011) (Permenkes) No. 1109/Menkes/Per/V1/2011 Petunjuk Teknis Jaminan Pelayanan Pengobatan Thalassaemia. Jakarta: Menteri Kesehatan Republik Indonesia.
- [5] Prasetyo, Eko. (2014). *Data Mining - Mengolah Data Menjadi Informasi menggunakan Matlab*. Yogyakarta: CV. Andi Offset.

- [6] Rejeki, Dwi Sarwani Sri. dkk (2014). *Model Prediksi Kebutuhan Darah untuk Penderita Talasemia Mayor*. Dalam Jurnal Kesehatan Masyarakat Nasional Vol. 8, No. 7, Februari 2014. [Online]. Tersedia: <http://download.portalgaruda.org/article.php?article=269759&val=7113&title=Model%20Prediksi%20Kebutuhan%20Darah%20untuk%20Penderita%20Talasemia%20Mayor>. [30 Agustus 2016]
- [7] Sani Susanto, D. S. (2010). *Pengantar Data Mining*. Yogyakarta: CV. Andi Offset.
- [8] Santosa, B. (2007). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.
- [9] Santoso, S. (2010). *Statistik Multivariat*. Jakarta: Elex Media Komputindo.
- [10] Sembiring, S.P.K. (2010). *Thalasemia*. Medan: MorphostLab E-Book Press. [Online]. Tersedia : <http://www.morphostlab.com> [22 April 2016]
- [11] Tahta, Budi, dan Ali. 2013. *Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya dalam Custer Data (Studi kasus: Problem Kerja Praktek Jurusan Teknik Industri ITS)*. Jurnal Skripsi Teknik Industri ITS. Surabaya: Teknik Industri, Institut Teknologi Sepuluh November (ITS).
- [12] Witten, Ian H. dan Frank, Eibe. 2005. *Data Mining Practical Machine Learning Tools and Techniques*, Second Edition. Morgan Kaufmann, San Fransisco
- [13] Wu, X. and Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. London: CRC Press Taylor & Francis Group.

BIODATA PENULIS



Heni Sulastri, M.T dilahirkan di Purwakarta, 01 November 1986. Sekolah Dasar dilaksanakan di SD Warnasari, Darangdan Purwakarta, SMP Negeri 1 Bojong, SMA Negeri 2 Purwakarta. Pendidikan S-1 (Teknik Informatika Univeritas Siliwangi) lulus tahun 2009 dan S-2 (Sistem Informasi ITB) lulus tahun 2017.



Acep Irham Gufroni, S.Kom., M.Eng. lahir di Tasikmalaya, 14 Maret 1985. menyelesaikan pendidikan tingkat atas di SMA Negeri 1 Tasikmalaya tahun 2003. Pendidikan S-1 (Informatika UII) lulus Tahun 2008 dan S-2 (Teknologi Informasi UGM) lulus tahun 2010.