# CASE BASE REASONING UNTUK MENDIAGNOSIS PENYAKIT HIPERTENSI MENGGUNAKAN METODE INDEXING DENSITY BASED SPATIAL CLUSTERING APPLICATION WITH NOISE (DBSCAN)

**[1]Herdiesel Santoso, [2]Andri Syafrianto**

*[1,2]Program Studi Sistem Informasi STMIK El Rahma, Yogyakarta*

*E-mail: [1]herdiesel.santoso@stmikelrahma.ac.id , [2]andrisyafrianto@stmikelrahma.ac.id*

***Abstract****. Hypertension is one of the health problems priority in the world because of the increasing of life expectancy and an unhealthy lifestyle. Many people with hypertension are unreachable and undiagnosed by a health worker and they do not do treatment according to the health recommendation. The Case-Based Reasoning (CBR) Method can be applied to solve the new cases in diagnosed hypertension using the answer or experience from the old case by comparing the new case and the old case. In order to do not use all the basic case data for finding a similar case, it makes an indexing process is needed. The DBSCAN algorithm implementation as indexing method is expected to improve the time and memory efficiency in CBR, especially during the retrieval process. The result of the CBR test with the cluster-indexing has a better accuracy and time process than the non-indexing CBR. The minimum parameter points and epsilon that has been chosen for clustering on hypertension data case is the combination of epsilon score 9 and minimum points score 3 with the silhoutte coefficient score 0.240 and average cluster time 0.541 seconds. The Minkowski Distance method has better accuracy than the Euclidean Distance method because by the threshold score ≥ 0.9 the CBR system with the Minkowski distance method is able to diagnose the disease with 100 % accuracy and the average best retrieval time, it is 0.0586 second.*

***Keywords: case based of hypertension reasoning, indexing, clustering, DBSCAN.***

***Abstrak.*** *Hipertensi menjadi salah satu prioritas masalah kesehatan di dunia karena peningkatan angka harapan hidup dan gaya hidup yang tidak sehat. Banyak penderita hipertensi yang tidak terjangkau dan terdiagnosis oleh tenaga kesehatan serta tidak menjalani pengobatan sesuai anjuran kesehatan. Metode Case-Based Reasoning (CBR) dapat diaplikasikan untuk menyelesaikan masalah baru dalam diagnosis penyakit hipertensi menggunakan jawaban atau pengalaman dari masalah lama dengan membandingkan kasus baru dengan kasus lama. Supaya proses pencarian kasus yang mirip tidak perlu melibatkan seluruh data pada basis kasus,maka diperlukan proses indexing. Implementasi algoritme DBSCAN sebagai metode indexing diharapkan dapat meningkatkan efisiensi waktu dan memori pada CBR khususnya ketika proses retrival. Hasil pengujian CBR dengan cluster-indexing memiliki akurasi dan waktu proses yang lebih baik dari pada CBR non-indexing. Parameter minimum points dan epsilon yang dipilih untuk melakukan clustering pada data kasus penyakit hipertensi adalah kombinasi epsilon 9 dan minimum points 3 dengan nilai silhoutte coeffisien 0.240 dan waktu klaster rata-rata 0.541 detik. Metode minkowski distance memiliki akurasi yang lebih baik dari pada metode euclidean distance, karena dengan threshold ≥ 0.9 sistem CBR dengan metode minkowski distance mampu mendignosis penyakit dengan akurasi 100% dan waktu retrieve rata-rata terbaik yaitu 0.0586 detik.*

***Kata kunci: hipertensi penalaran berbasis kasus, indexing, clustering, DBSCAN.***

## 1. Pendahuluan

Hipertensi atau yang lebih dikenal dengan penyakit darah tinggi adalah suatu keadaan dimana seseorang mengalami peningkatan tekanan darah diatas normal. Seringkali penderita hipertensi tidak merasakan sakit dan jarang menampakkan gejala yang jelas dalam waktu yang lama. Tanpa disadari penderita mengalami komplikasi pada organ vitalnya. Gejala-gejala akibat hipertensi seperti pusing, gangguan penglihatan, dan sakit kepala sering muncul disaat hipertensi sudah berada ditahap lanjut dan pada tekanan darah tertentu, karena itu hipertensi sering disebut sebagai "pembunuh secara diam-diam (Triyanto, 2014). Jika tidak segera ditangani, dalam jangka panjang peningkatan tekanan darah yang berlangsung kronik akan menyebabkan peningkatan resiko kejadian penyakit lainnya, seperti kardiovaskuler, serebrovaskuler dan renovaskuler (Firmansyah, Lukman, & Mambangsari, 2017).

Penderita penyakit hipertensi masih cukup tinggi dan bahkan cenderung meningkat karena gaya hidup yang tidak sehat seperti kelebihan berat badan, merokok, kurang berolahraga, diet yang tidak sehat, stres dan sebagainya (Dhianingtyas & Hendrati, 2006). Selain itu mahalnya biaya pengobatan, kurangnya sarana dan prasarana penanggulangan hipertensi terutama untuk melakukan deteksi dini hipertensi menjadi faktor masih banyak penderita hipertensi yang tidak terjangkau dan terdiagnosis oleh tenaga kesehatan (Tedjasukmana, 2012). Berdasarkan penelitian yang dilakukan di 167 negara berkembang menunjukkan bahwa sebanyak 45% dari tenaga kesehatan profesional tidak dilatih untuk mengelola permasalahan hipertensi dan sebanyak 61% dari sejumlah negara tersebut tidak memiliki panduan nasional tentang penatalaksanaan hipertensi (Whitworth, 2003) .

*Case-Based Reasoning* (CBR) telah diaplikasikan pada banyak bidang, terutama dibidang kedokteran (Mulyana & Hartati, 2009). CBR bekerja dengan meniru kemampuan manusia, yaitu menyelesaikan masalah baru menggunakan jawaban atau pengalaman dari masalah lama dengan membandingkan kasus baru dengan kasus lama. Jika kasus baru tersebut mempunyai kemiripan dengan kasus lama maka CBR akan memberikan jawaban kasus lama untuk kasus baru tersebut. Jika tidak ada yang cocok maka CBR akan melakukan adaptasi, dengan cara memasukkan kasus baru tersebut ke dalam penyimpanan kasus (Pal & Shiu, 2004). Sehingga, semakin lama data kasus dalam basis kasus akan semakin banyak dan secara tidak langsung pengetahuan CBR akan bertambah.

Jika kasus lama yang ada pada basis kasus memiliki jumlah yang banyak, proses untuk menemukan kasus yang relevan akan memerlukan waktu yang lama, karena sistem harus menghitung nilai kemiripan kasus baru terhadap semua kasus lama yang ada di basis kasus. Supaya proses pencarian kasus yang mirip tidak perlu melibatkan seluruh data pada basis kasus, tetapi cukup pada beberapa kasus terdekat, diperlukan proses *indexing*. *Indexing* akan bekerja dengan cara mengelompokkan kasus-kasus yang ada berdasarkan fitur yang ditentukan. *Clustering* merupakan teknik eksplorasi yang kuat untuk mengekstraksi pengetahuan dari sekumpulan data. *Clustering* dapat mengelompokan sejumlah data yang tidak berlabel berdasarakan kemiripan (*similarity*) dan ketidakmiripan (*dissimilarity*) ke dalam kelompok yang disebut *cluster*, sehingga dalam setiap *cluster* akan berisi data yang semirip mungkin (Witten & Frank, 2005).

https://doi.org/10.29313/ethos.v7i1.4206

Penelitian ini bertujuan membangun sistem CBR untuk membantu diagnosis penyakit hipertensi. Metode *indexing* yang diusulkan menggunakan algoritme *clustering* berbasis densitas yaitu *Density Based Spatial Clustering Application with Noise* (DBSCAN). Implementasi algoritme DBSCAN sebagai metode *indexing* diharapkan dapat meningkatkan efisiensi waktu dan memori pada CBR khususnya ketika proses *retrival*. Proses perhitungan similaritas membandingkan metode *euclidean distance* dan *minkowski distance similarity*. Sistem ini diharapkan mampu mendeteksi jenis hipertensi dan memberikan solusi untuk penyakit hipertensi secara cepat dan tepat berdasarkan kemirirpan kasus pada kasus-kasus terdahulu.

## 2. Metode Penelitian

### *State Of The Art*

Penelitian yang berfokus pada case base reasoning telah banyak dilakukan dengan berbagai macam metode retrival, indexing, dan berbagai macam bentukan kasus. Penelitian labellapansa dkk menerapkan CBR pada kasus penyakit skizofrenia, menggunakan metode weighted Minkowski distance dengan lambda (r) bernilai 1, 2 dan 3 untuk menghitung ukuran jarak kedekatan antar kasus. Nilai akurasi tertinggi didapatkan ketika lambda (r) pada metode weighted Minkowski distance bernilai 3 (Labellapansa, Efendi, Yulianti, & Evizal, 2016).

Penelitian yang telah menerapkan metode *indexing* di antaranya dilakukan oleh Mohsin dkk membahas tentang penggunaan metode hashing untuk melakukan proses indexing pada *case base* menggunakan data kasus dari operasional bendungan Timah Tasoh. Hasil dari penelitian ini menunjukkan bahwa proses retrieval dengan indexing menggunakan metode hashing memberikan hasil yang lebih baik daripada penggunaan metode CBR secara konvensional (Mohsin, Manaf, Norwawi, & Wahab, 2011). Penelitan Riswawan dan Hartati, menggunakan metode *backpropagation* untuk metode *indexing* dengan kasus untuk diagnosis penyakit THT. Sedangkan proses perhitungan *similarity* dengan menggunakan *cosine coeffecient*. Hasil dari penelitian bahwa dengan penggunaan metode *backpropagation* pada proses indexing dapat membantu sistem dalam melakukan *retrieval* karena dengan menggunakan *backpropagation*, pencarian nilai *similarity* cukup dilakukan terhadap kasus yang memiliki indeks yang sama dengan kasus baru (Rismawan & Hartati, 2012). Penelitian yang dilakukan Kim dan Han membandingkan hasil indexing dengan metode induktif, *self-organizing maps* (SOM) dan *learning vector quantization* (LVQ) pada CBR, sedangkan proses perhitungan similarity menggunakan metode *nearest neighbor* dengan pengukuran jarak *euclidian*. Hasil penelitian menunjukan bahwa CBR dengan proses indexing memiliki performa yang lebih tinggi dari CBR tanpa indexing, dan perbandingan dari metode induksi, SOM dan LVQ memperoleh hasil bahwa metode yang terbaik adalah LVQ, kemudian SOM dan terakhir adalah metode induksi (Kim & Han, 2001). Metode berbasis jaringan saraf tiruan seperti SOM, LVQ dan *backpropagation* memiliki kekurangan yaitu dalam proses pelatihan, memerlukan waktu yang cukup lama karena harus mencoba parameter pelatihan satu per satu untuk memperoleh jaringan yang terbaik. Metode ini juga memiliki toleransi yang rendah terhadap data yang mengandung *noise* dan *outlier*.

DBSCAN merupakan algoritme *clustering* yang masuk dalam kategori

*density-based clustering*, yaitu proses pembentukan klaster dilakukan berdasarkan tingkat kedekatan/kerapatan jarak antar obyek dalam dataset tersebut (Ester, Kriegel, Sander, & Xu, 2010). DBSCAN memiliki kelebihan yaitu dapat menangani data dalam jumlah yang besar dengan waktu yang cepat, memiliki toleransi terhadap data yang mengandung noise dan outlier, dapat menangani data berdimensi tinggi, serta tidak perlu mendefinisikan jumlah klaster yang akan terbentuk (Parimala, Lopez, & Senthilkumar, 2011). Berdasarkan hasil eksperimen penelitian, algoritme DBSCAN dapat diimplementasikan untuk mengelompokkan sejumlah data dengan hasil yang cukup bagus dengan cukup tingginya nilai *Silhouette Coefficient* pada proses pengujian (Furqon & Muflikhah, 2016).

## Akusis Pengetahuan

Data primer diperoleh dari catatan rekam medis pasien rawat inap terdiagnosis hipertensi periode tahun 2013-2017 yang diperoleh dari RS PKU Muhammadiyah Yogyakarta dan hasil wawancara dengan dokter spesialis penyakit dalam. Data penyakit yang dipakai adalah penyakit hipertensi yang terdiri dari normal, pre-hipertensi, hipertensi tingkat-1, hipertensi tingkat-2, hipertensi urgensi dan hipertensi emergensi, yang masing-masing disertai dengan rekomendasi terapi. Data fitur simbolik adalah data gejala yang terdiri dari 10 gejala yaitu jantung berdebar, mual, pusing, lemas, sakit kepala, nyeri dada, sesak napas, mimisan, pandangan kabur dan hilang kesadaran, serta data riwayat penyakit yang terdiri dari 5 riwayat penyakit yaitu hipertensi, diabetes mellitus, stroke, jantung dan ginjal. Data fitur numerik untuk penyakit hipertensi adalah usia, jenis kelamin, nilai tekanan darah sistolik (TDS), nilai tekanan darah diastolic (TDD), jumlah nadi dan jumlah napas.

## Representasi Kasus

*Model frame* digunakan untuk merepresentasikan kasus-kasus yang didapat melaui akusis pengetahuan. Kasus direpresentasikan dalam bentuk kumpulan fitur-fitur yang menjadi ciri kasus tersebut dan solusi untuk menangani kasus tersebut. Dalam setiap fitur kasus, baik itu faktor gejala maupun faktor resiko, memiliki bobot yang menunjukan tingkat kepentingan terhadap penyakit yang diderita pasien. Pada data kasus penyakit hipertensi, bobot fitur antara 0 sampai 10. Representasi kasus penyakit hipertensi yang telah ditambahkan pengetahuan baru yang berasal dari nilai pusat klaster di tunjukan pada Tabel 1.

**Tabel 1** Representasi kasus penyakit hipertensi

| Fitur | Nilai |
|---|---|
| Usia | Numerik |
| Jenis kelamin | Laki-laki = 1 dan Perempuan = 0 |
| Tekanan darah sistolik (TDS)/mmHg | Numerik |
| Tekanan darah diastolik (TDD)/mmHg | Numerik |
| Jumlah nadi/menit | Numerik |
| Jumlah napas/menit | Numerik |
| **Gejala** | |
| Jantung berdebar | Ya = 1 dan Tidak = 0 |
| Mual | Ya = 1 dan Tidak = 0 |
| Pusing | Ya = 1 dan Tidak = 0 |
| Lemas | Ya = 1 dan Tidak = 0 |
| Sakit kepala | Ya = 1 dan Tidak = 0 |
| Nyeri dada | Ya = 1 dan Tidak = 0 |
| Sesak napas | Ya = 1 dan Tidak = 0 |
| **Riwayat penyakit** | |
| Hipertensi | Ya = 1 dan Tidak = 0 |
| Diabetes mellitus | Ya = 1 dan Tidak = 0 |
| Stroke | Ya = 1 dan Tidak = 0 |
| Jantung (HHD) | Ya = 1 dan Tidak = 0 |

https://doi.org/10.29313/ethos.v7i1.4206

| | |
|---|---|
| **Diagnosis**<br>P01<br>P02<br>P03<br>P04<br>P05 | <br>Pre-hipertensi<br>Hipertensi tingkat-1<br>Hipertensi tingkat-2<br>Hipertensi urgensi<br>Hipertensi emergensi |
| **Klaster** | Numerik |
| **Rekomendasi** | Terapi non-farmakologi<br>Terapi farmakologi |
| **Solusi tindakan** | Modifikasi gaya hidup<br>Pemberian obat hipertensi |

Jika ada kasus baru pembobotan terhadap fitur penyakit dibagi menjadi dua kategori yaitu Tidak dan Ya. Nilai untuk masing-masing kategori adalah 0 untuk Tidak dan 1 untuk Ya. Setelah kasus-kasus lama yang ada di basis kasus diklasterkan, data kasus lama direpresentasikan kembali dengan menambahkan pengetahuan baru yang berasal nilai pusat klaster (cluster centroids).

## Indexing dengan Clustering

*Indeks* pada CBR merupakan struktur data yang terletak dalam memori utama dan dapat mempercepat pencarian, sehingga CBR tidak perlu mencari tiap kasus yang ada dalam basis kasus yang tentunya akan lambat. Metode *indexing* yang digunakan dalam sistem ini menggunakan metode *clustering* yaitu *Density Based Spatial Clustering Application with Noise* (DBSCAN). DBSCAN digunakan untuk mengelompokan data kasus lama ke dalam kelompok-kelompok berdasarakan kemiripan (*similarity*) dan ketidakmiripan (*dissimilarity*), sehingga dalam setiap kelompok akan berisi data yang semirip mungkin.

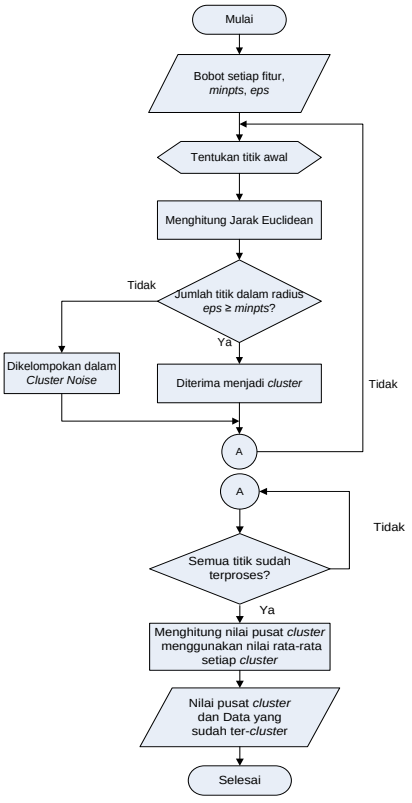
Normalisasi Data

Normalisasi ini bertujuan untuk mendapatkan data dengan ukuran yang lebih kecil yang mewakili data yang asli tanpa kehilangan karakteristik sendirinya. Data yang dilakukan normalisasi adalah data fitur usia, nilai tekanan darah sistolik (TDS), nilai tekanan darah diastolic (TDD), jumlah nadi dan jumlah napas karena mempunyai nilai rentan yang cukup besar. Normalisasi menggunakan metode *Min Max Normalization*. *Min Max Normalization* membutuhkan nilai Minimum dan Maksimum fitur usia. Persamaan 1 merupakan rumus *Min Max Normalization* (Han & Kamber, 2006) .

$$v' = \frac{v - minA}{maxA - minA} \qquad (1)$$

*Density Based Spatial Clustering Application with Noise (DBSCAN)*

*Density Based Spatial Clustering Application with Noise* (DBSCAN) adalah salah satu algoritma *clustering density-based*. Algoritma memperluas wilayah dengan kepadatan yang tinggi ke dalam klaster dan menempatkan klaster *irregular* pada basis data spasial dengan *noise*. Rancangan *clustering* dengan menggunakan metode DBSCAN ditunjukan pada diagaram alir Gambar 2.




**Gambar 2** Rancangan *clustering* menggunakan metode DBSCAN

DBSCAN memiliki 2 parameter yaitu *Eps* (radius maksimum dari neighborhood) dan *MinPts* (jumlah minimum titik dalam Eps-neighborhood dari suatu titik). Penjelasan diagram alir Gambar 2 adalah sebagai berikut:

1. Melakukan inisialisasi :
    a. bobot setiap fitur dalam basis kasus sebagai *input* dari DBSCAN,
    b. radius maksimum dari *neighborhood* (*Eps*) dan jumlah minimum titik dalam *Eps-neighborhood* dari suatu titik (*MinPts*) sebagai parameter DBSCAN.
2. Menentukan satu data kasus sebagai titik awal atau p secara acak .
3. Untuk setiap data kasus dalam basis kasus hitung nilai *Eps* atau semua jarak yang *density reachable* terhadap p menggunakan persamaan (2).

$$D_{ij=}\sqrt{\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)^2} \quad (2)$$

4. Jika jumlah data kasus yang memenuhi *Eps* lebih dari *MinPts* maka p adalah *core point* dan terbentuk satu klaster.
5. Jika tidak ada data kasus yang yang *density reachable* terhadap p atau jumlah data kasus yang memenuhi *Eps* kurang dari *MinPts* maka p adalah *Noise*.
6. Mengulangi langkah 3 – 5 hingga semua data basis kasus diproses.
7. Menghitung nilai pusat klaster (*cluster centroid*) menggunakan nilai rata-rata untuk masing-masing kelompok *cluster*.
8. *Output* dari data basis kasus yang sudah terklaster dan nilai rata-rata yang digunakan sebagai nilai pusat klaster.

*Metode Evaluasi Klaster*

Metode evaluasi yang akan digunakan pada sistem ini adalah metode *silhouette coefficient*. Metode ini berfungsi untuk menguji kualitas dari klaster yang dihasilkan. Metode ini merupakan metode validasi klaster yang menggabungkan metode *cohesion* dan *separation*. Untuk menghitung nilai *silhoutte coefisient* diperlukan jarak antar data dengan menggunakan rumus *euclidean distance*. Tahapan untuk menghitung nilai *silhoutte coeffisien* adalah sebagai berikut (Rousseeuw, 1987):

1. Untuk setiap objek i, hitung rata-rata jarak dari objek i dengan seluruh objek yang berada dalam satu *cluster*. Akan didapatkan nilai rata-rata yang disebut *ai*.
2. Untuk setiap objek i, hitung rata-rata jarak dari objek i dengan objek yang berada di klaster lainnya. Dari semua jarak rata-rata tersebut ambil nilai yang paling kecil, nilai tersebut disebut dengan *bi*.
3. Selanjutnya hitung *silhoutte coeffisien* dengan persamaan 3.

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (3)$$

4. Pada akhirnya, dilakukan pencarian rata-rata dari lebar *silhouette* untuk kelas C, yang selanjutnya digunakan untuk mendapatkan *silhouette coefficient global* yang merupakan rata-rata nilai dari rata-rata *silhouette* semua klaster.

$$\mathrm{S_k} = \frac{1}{\mathrm{n_k}} \sum_{\mathrm{i} \in \mathrm{I_k}} \mathrm{s(i)} \quad (4)$$

$$\mathrm{C} = \frac{1}{\mathrm{K}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{S_k} \quad (5)$$

Nilai *silhoutte coeffisien* dapat bervariasi antara -1 hingga 1. Hasil *clustering* dikatakan baik jika nilai *silhoutte coeffisien* bernilai positif (*ai* < *bi*) dan *ai* mendekati 0, sehingga akan menghasilkan nilai *silhoutte coeffisien* yang maksimum yaitu 1. Maka dapat dikatakan objek i sudah berada dalam klaster yang tepat. Jika nilai *silhoutte coeffisien* sama dengan 0 maka objek i berada di perbatasan antara dua *cluster*.

https://doi.org/10.29313/ethos.v7i1.4206

Tetapi, jika *silhoutte coeffisien* sama dengan -1 artinya objek i lebih tepat dimasukan ke dalam klaster yang lain. *Nilai rata-rata silhoutte coeffisien* dari tiap objek dalam suatu klaster adalah suatu ukuran yang menunjukan seberapa ketat data dikelompokan dalam klaster tersebut.

*Proses Retrieve dan Reuse*

Pengukuran similaritas atau *retrieval* merupakan suatu proses menemukan kasus-kasus sebelumnya yang disimpan di basis kasus yang kemudian digunakan kembali untuk mendapatkan solusi dari permasalahan baru. Sistem CBR dengan *cluster-indexing* tidak perlu menghitung nilai kemiripan kasus baru terhadap semua kasus yang ada pada basis kasus tetapi cukup menghitung nilai kemiripan terhadap kasus yang berada pada kelompok yang sama.

*Penentuan Klaster Terdekat*

Pada saat melakukan proses pencarian kasus yang mirip dengan kasus yang baru, sistem CBR akan melakukan pencarian klaster yang paling relevan dengan kasus baru dengan cara menghitung kemiripan gejala kasus lama dengan nilai pusat klaster. Proses perhitungan kemiripan dengan cara membandingkan jarak menggunakan metode *cosine coefficient.* Jika diberikan 2 buah vektor X dan Y, maka nilai kesamaannya dapat dicari dengan persamaan 6 (Zhu, Wu, Xiong, & Xia, 2011).

$$Cos\ (X,Y) = \frac{\langle X,Y\rangle}{||X||||Y||} \quad (6)$$

dimana "‹›" menunjukkan perkalian dari vector X dan Y, dan "|| ||" menunjukkan norm pada masing-masing vektor. Untuk vektor dengan elemen elemen non-negatif, nilai kemiripan *cosinus* selalu terletak antara 0 dan 1, dimana 1 menunjukkan kedua vektor tersebut benar-benar sama dan nilai 0 menunjukkan sebaliknya.

*Pengukuran Similaritas*

Penelitian ini menggunakan metode *nearest neighbor* untuk proses *retrieval*. Konsep dasar dari *nearest neighbor* adalah suatu pendekatan untuk mencari kasus dengan menghitung *similarity metric*, yaitu kedekatan antara kasus baru dengan kasus lama berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Terdapat 2 (dua) macam pengukuran similaritas, yaitu similaritas lokal dan similaritas global. Similaritas lokal adalah pengukuran kedekatan pada level fitur, sedangkan similaritas global adalah pengukuran kedekatan pada level objek (kasus).

Similaritas lokal yang digunakan dalam penelitian ini dibedakan menjadi 2 (dua) jenis yaitu numerik dan simbolik. Data yang bersifat numerik akan dihitung menggunakan persamaan 7 (Pal & Shiu, 2004).

$$f(S_i,T_i) = 1 - \frac{|S_i,T_i|}{|f_{max}-f_{min}|} \quad (7)$$

Keterangan: $f(S_i,T_i)$ adalah kesamaan fitur ke-$i$ dari kasus S (*source*) dan kasus T (*target*), $S_i$adalah nilai fitur ke-$i$ dari kasus lama (*source*), $T_i$ adalah nilai fitur ke-$i$ dari kasus baru (*target*), $f_{max}$ adalah nilai maksimum fitur ke- $i$ pada basis kasus dan $f_{min}$ adalah nilai minimum fitur ke- $i$ pada basis kasus. Sedangkan data yang bersifat simbolik akan dihitung menggunakan persamaan 8.

$$f(S_i,T_i) = \begin{cases} 0, & jika\ S_i \neq T_i \\ 1, & jika\ S_i = T_i \end{cases} \quad (8)$$

Keterangan: $f(S_i,T_i)$ adalah kesamaan fitur ke-$i$ dari kasus S (*source*) dan kasus T (*target*), $S_i$adalah nilai fitur ke-$i$ dari kasus lama (*source*) dan $T_i$ adalah nilai fitur ke-$i$ dari kasus baru (*target*).

Similaritas global digunakan untuk menghitung keserupaan antar masalah baru dengan kasus yang tersimpan dalam basis kasus.

$$Sim\ (S,T) = \left(\frac{\sum_{i=1}^{n} {w_i}^r \times |f(S_i,T_i)|^r}{\sum_{i=1}^{n} {w_i}^r}\right)^{1/r} \quad (9)$$

Keterangan: $Sim(S,T)$ adalah nilai similaritas antara kasus lama (S) dan kasus baru (T), *fi (Si,Ti)* adalah kesamaan fitur ke-i dari *source case* dan *target case*,kesamaan fitur ke-i dari *source case* dan *target case*, $n$ adalah jumlah fitur pada masing-masing kasus, $i$ adalah fitur individu, antara 1 s/d n, $w_i$ adalah bobot yang diberikan pada fitur ke-$i$, dan $r$ adalah faktor *minkowski* (*integer* positif). Nilai $r$ adalah bilangan positif ≥ 1. Pada penelitian ini akan menggunakan $r$ sama dengan 2 dan $r$ sama dengan 3. Jika nilai $r$ sama dengan 2 dikenal dengan *euclidean distance* (Merigó & Casanovas, 2011) dan jika nilai $r$ sama dengan 3 dikenal dengan *minkowski distance similarity* (Núñez et al., 2004).

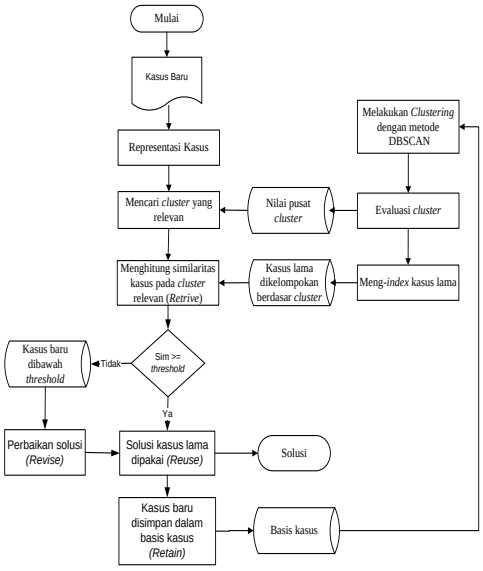
*Arsitektur Sistem*

Gambaran umum sistem yang dibangun mengikuti arsitektur CBR *cycle* yaitu *retrieve*, *reuse*, *revise* dan *retain* dengan beberapa penyesuaian berdasarkan dengan sistem yang dikembangkan. Gambar 3 menunjukan gambaran umum sistem yang dibangun. Proses diagnosis menggunakan CBR dengan *cluster-indexing* dapat dijelaskan sebagai berikut:

1. Jika terdapat kasus baru, sistem menginisialisasi gejala yang dialami oleh pasien dan merepresentasikanya sebagai kasus baru.
2. Sistem akan melakukan pencarian klaster yang paling relevan dengan cara menghitung kemiripan gejala kasus baru dengan nilai pusat klaster. Proses perhitungan kemiripan dengan cara membandingkan jarak antara kasus baru dengan nilai pusat klaster menggunakan metode *cosine coefficient* persaman 6.
3. Setelah memperoleh *indeks* atau klaster yang relevan dengan kasus baru, kemudian dihitung nilai similaritas antara kasus baru tersebut dengan source case atau kasus-kasus yang ada di basis kasus yang berada pada klaster yang sama (*retrieve*). Teknik untuk mencari nilai kemiripan antara kasus lama (*old case*) dengan permasalahan baru (*problem case*) menggunakan pengukuran similaritas persamaan 9. Data yang bersifat numerik akan dihitung menggunakan persamaan 7 dan simbolik akan dihitung menggunakan persamaan 8. Nilai r yang digunakan adalah r sama dengan 2 untuk metode *euclidean distance similarity* dan r sama dengan 3 untuk metode *minkowski distance similarity*. Nilai similaritas berada antara 0 sampai dengan 1.
4. Nilai *threshold* similaritas yang digunakan adalah 0.9 yang berarti jika similaritas tertinggi lebih besar dari threshold dan mendekati 1 hal ini menandakan bahwa kasus baru tersebut memiliki kemiripan yang sama persis dengan kasus lama maka solusi dari *source case* akan diberikan kepada user (*reuse*).
5. Jika nilai similaritas semakin kecil atau dibawah *threshold*, menandakan bahwa kasus baru tersebut semakin tidak mirip dengan kasus lama yang ada di *case base*. Kasus tersebut akan disimpan dalam basis data sebagai kasus revisi yang nantinya kasus yang dibawah *threshold* tersebut akan dilakukan penyesuaian dari solusi kasus-kasus sebelumnya oleh pakar (*revise*).

https://doi.org/10.29313/ethos.v7i1.4206

6. Kasus baru kemudian disimpan ke dalam case base dengan mempertimbangkan nilai pusat klaster untuk menjadi pengetahuan baru (*retain*).
7. Proses *clustering* kasus-kasus lama yang ada di basis kasus dengan menggunakan algoritme DBSCAN. Penetuan jarak *euclidean* mengunakan persamaan 2. Hasil dari clustering adalah nilai pusat klaster (*cluster centroids*) yang digunakan untuk proses *indexing*.
8. Evaluasi klaster, nilai pusat klaster (cluster centroids) tersebut dievaluasi menggunakan metode *silhoutte coeffisien* menggunakan persamaan 5. Nilai pusat klaster disimpan di dalam basis data.
9. Meng-Index kasus lama, setelah kasus-kasus lama yang ada di basis kasus diklasterkan, data kasus lama diperbaharui kembali dengan menambahkan pengetahuan baru yang berasal nilai pusat klaster (*cluster centroids*), kemudian disimpan sebagai basis kasus yang sudah terindeks.




**Gambar 3.** Arsitektur sistem

Pengujian Sistem

Pengujian dilakukan dengan menerapkan permasalahan baru yaitu 20 kasus sebagai data uji. Hasil dari sistem kemudian dibandingkan dengan data yang tertera pada data rekam medis. Tahapan-tahapan dari pengujian sistem seperti terlihat pada Gambar 4.

## 3. Hasil Dan Pembahasan

Analisis kemampuan sistem diagnosis penyakit terbagi kedalam dua skenario. Skenario pertama diagnosis sistem dengan menggunakan CBR *non-indexing*, skenario kedua diagnosis CBR dengan *indexing* menggunakan metode DBSCAN. Proses pencarian klaster yang relevan pada CBR dengan *indexing* menggunakan metode *cosine similarity* dan proses perhitungan similaritas unuk ketiga skenario menggunakan *minkowski distance similarity*, dan *euclidean distance similarity*. Pengujian dilakukan dengan menerapkan permasalahan baru yaitu 20 kasus sebagai data uji untuk masing-masing data kasus.

Skenario 1

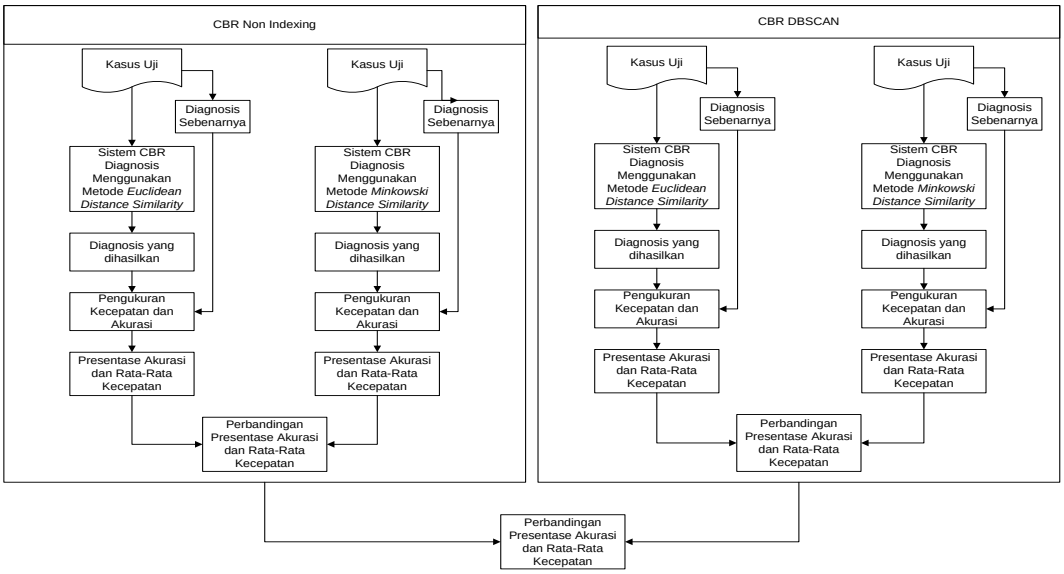
Skenario 1 merupakan pengujian sistem dengan menggunakan CBR *non-indexing*. Rekapitulasi hasil pengujian dengan kasus data uji penyakit hipertensi diperlihatkan pada Tabel 2.

**Tabel 2.** Hasil pengujian data kasus hipertensi metode *non-indexing*

| Metode | Euclidean Distance | Minkowski Distance |
|---|---|---|
| Akurasi *threshold* ≥ 0.9 | 75% | 100% |
| Waktu *retrieve* rata-rata (detik) | 0.069 | 0.066 |

Skenario 2

Skenario 2 merupakan pengujian sistem diagnosis CBR menggunakan metode *indexing* DBSCAN. Parameter Metode DBSCAN membutuhkan 2 (dua) buah parameter, yaitu *minimum points* dan *epsilon*.




**Gambar 4.** Skema pengujian sistem CBR dengan *clustering* sebagai *indexing*

Nilai parameter dikatakan optimal apabila dapat menghasilkan nilai *silhoutte coeffisien* dan akurasi yang lebih besar. Proses penentuan parameter DBSCAN yang optimal dilakukan dengan cara melakukan *clustering* terhadap setiap *dataset* menggunakan beberapa kombinasi *minimum points* dan *epsilon*. Pada percobaan awal yang telah dilakukan, nilai *epsilon* yang terlalu kecil atau nilai *minimum points* yang terlalu besar akan menghasilkan banyak *noise*, bahkan dapat menyebabkan seluruh data dikelompokan sebagai *noise*.

Kandidat nilai parameter DBSCAN optimal diperoleh untuk setiap *epsilon* dengan nilai *silhoutte coeffisien* dari hasil *clustering*. Kemudian setiap kombinasi parameter tersebut digunakan untuk menghitung waktu dan akurasi proses *retrieve* CBR. Setiap percobaan dibandingkan waktu tercepat dan nilai akurasinya, nilai tertinggi dipilih sebagai parameter DBSCAN optimal.

**Tabel 3.** Hasil pengujian parameter DBSCAN data kasus hipertensi

| Epsilon | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|
| *MinPoints* | 3 | 3 | 3 | 3 | 3 | 3 |
| Jumlah *cluster* | 4 | 3 | 3 | 4 | 3 | 3 |
| Jumlah Noise | 62 | 40 | 30 | 20 | 15 | 14 |
| *Silhoutte Coeffisien* | 0.0522 | 0.1999 | 0.2300 | 0.2400 | 0.3122 | 0.3221 |
| Waktu Cluster (detik) | 0.5700 | 0.5443 | 0.5466 | 0.5441 | 0.5554 | 0.5700 |
| **Metode Euclidean Distance** | | | | | | |
| Akurasi *threshold* ≥ 0.9 | 70% | 70% | 65% | 65% | 65% | 65% |
| Waktu rata-rata Retrieve (detik) | 0.05540 | 0.06007 | 0.06009 | 0.06029 | 0.08001 | 0.07996 |
| **Metode Minkowski Distance** | | | | | | |
| Akurasi *threshold* ≥ 0.9 | 85% | 90% | 100% | 100% | 100% | 100% |
| Waktu rata-rata Retrieve (detik) | 0.05558 | 0.05591 | 0.06004 | 0.05586 | 0.07001 | 0.07996 |

Hasil pengukuran akurasi pada data kasus hipertensi dapat disimpulkan bahwa nilai kombinasi *minimum points* dan *epsilon* yang berbeda dapat menghasilkan jumlah klaster dan jumlah *noises* berbeda. Pada metode Metode *Euclidean Distance*, akurasi tertinggi

didapatkan dari kombinasi *epsilon* dan *minimum points* yaitu {{6;3}, {7;3}}, dengan nilai akurasi 70%. Hasil *clustering* tersebut memiliki nilai *silhoutte coeffisien,* yaitu {0.0520; 0.199}. Sedangkan untuk metode *minkowski distance*, akurasi tertinggi didapatkan dari kombinasi *epsilon* dan *minimum points distance* yaitu {{8;3}, {9;3}, {10;3}, {11;3}} yang menghasilkan nilai akurasi 100%. Kombinasi pada metode *minkowski distance* memiliki nilai *silhoutte coeffisien*, yaitu {0.230; 0.240; 0.250; 0.312; 0.321}. Nilai tersebut berarti setiap data sudah berada pada klaster yang tepat dan tidak ada kelas yang *overlap* pada kelas yang lainnya. Sehingga kedua kombinasi parameter DBSCAN dapat digunakan untuk memperoleh klaster yang optimal. Ketika proses pencarian indeks klaster yang relevan, *noise* dianggap sebagai klaster dan dimasukan dalam perhitungan similaritas. Sehingga jika kasus yang paling mirip terdapat pada klaster *noise*, solusi tetap akan diberikan. Parameter yang dipilih untuk melakukan *clustering* pada data kasus penyakit hipertensi adalah kombinasi *epsilon* 9 dan *minimum points* 3 dengan metode *minkowski distance*, karena memiliki akurasi 100%, nilai *silhoutte coeffisien* > 0 dan rata-rata waktu *retrieve* terbaik.

## 4. Kesimpulan dan Saran

### Kesimpulan

1. Hasil pengujian CBR dengan *cluster-indexing* memiliki akurasi sama baik dan waktu proses yang lebih baik dari pada CBR *non-indexing*.
2. Parameter *minimum points* dan *epsilon* yang dipilih untuk melakukan *clustering* adalah kombinasi *epsilon* 9 dan *minimum points* 3 dengan nilai *silhoutte coeffisien* 0.240 dan waktu klaster rata-rata 0.541 detik. Metode *minkowski distance* memiliki akurasi yang lebih baik dari pada metode *euclidean distance*, karena dengan *threshold* > 90% sistem CBR dengan metode *minkowski distance* mampu mendiagnosis penyakit dengan akurasi 100% dan waktu *retrieve* rata-rata terbaik yaitu 0.0586 detik.

### Saran

1. Peneitian selanjutnya perlu mempertimbangkan tingkat keyakinan pada permasalahan baru dan tingkat keyakinan pakar suatu kasus dalam melakukan perhitungan nilai similaritas karena perbedaan fitur-fitur yang ada dalam suatu kasus tertentu.
2. Perlu ditambahkan beberapa domain penyakit yang memiliki data kasus lebih banyak agar waktu *retrieve* kasus baru antara CBR *indexing* dan CBR non-*indexing* terlihat perbedaannya.

## Daftar Pustaka

Dhianingtyas, Y., & Hendrati, L. Y. 2006. Risiko Obesitas , Kebiasaan Merokok , dan Konsumsi Garam. *Media The Indonesian Journal of Public Health*, *2*(3), 2006.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. 2010. Density-Based Clustering Methods. *Comprehensive Chemometrics*, *2*, 635–654. https://doi.org/10.1016/B978-044452701-1.00067-3

Firmansyah, R. S., Lukman, M., & Mambangsari, C. W. 2017. Faktor-Faktor yang Berhubungan dengan Dukungan Keluarga dalam Pencegahan Primer Hipertensi Analysis of Factors Related to Support Families in Primary Prevention of Hypertension. *Jkp*, *5*, 197–213.

Furqon, M. T., & Muflikhah, L. 2016. Clustering the Potential Risk of Tsunami Using Density-Based Spatial Clustering of Application With Noise (Dbscan). *Journal of Enviromental Engineering and Sustainable Technology*, *3*(1), 1–8. https://doi.org/10.21776/ub.jeest.2016.003.01.1

Han, J., & Kamber, M. 2006. *Data mining concept and technique*. *Morgan Kaufmann* (2nd ed.). San Francisco: Morgan Kaufmann. https://doi.org/10.1017/CBO9781107415324.004

Kim, K. S., & Han, I. 2001. The cluster-indexing method for casebased reasoning using self-organizing maps and learning vector quantization for bond rating cases, Expert Systems with Applications. *Expert Systems with Applications*, *21*(3), 147–156. https://doi.org/10.1016/S0957-4174(01)00036-7

Labellapansa, A., Efendi, A., Yulianti, A., & Evizal, A. K. 2016. Lambda value analysis on Weighted Minkowski distance model in CBR of Schizophrenia type diagnosis. *2016 4th International Conference on Information and Communication Technology, ICoICT 2016*, *4*(c), 1–4. https://doi.org/10.1109/ICoICT.2016.7571898

Merigó, J. M., & Casanovas, M. 2011. A new minkowski distance based on induced aggregation operators. *International Journal of Computational Intelligence Systems*, *4*(2), 123–133. https://doi.org/10.1080/18756891.2011.9727769

Mohsin, M. F. M., Manaf, M., Norwawi, N. M., & Wahab, M. H. A. 2011. Faster Case Retrieval Using Hash Indexing Technique. *International Journal of Artificial Intelligence and Expert Systems (IJAE)*, *2*(2), 81–95.

Mulyana, S., & Hartati, S. 2009. Tinjauan Singkat Perkembangan Case – Based Reasoning. *semnasIF UPNVN Yogyakarta*, *2009*(semnasIF), 17–24.

Núñez, H., Sànchez-Marrè, M., Cortés, U., Comas, J., Martínez, M., Rodríguez-Roda, I., & Poch, M. (2004). A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. *Environmental Modelling and Software*, *19*(9), 809–819. https://doi.org/10.1016/j.envsoft.2003.03.003

Pal, S. K., & Shiu, S. C. 2004. *Foundations of Soft Case-Based Reasoning*. *John Willey and Sons, Inc.* Hoboken, New Jersey: John Willey and Sons, Inc. https://doi.org/10.1016/S0261-5614(03)00132-8

Parimala, M., Lopez, D., & Senthilkumar, N. C. 2011. A survey on density based clustering algorithms for mining large spatial databases. *International Journal of Advanced Science and Technology*, *31*(1), 59–66. https://doi.org/10.1002/14651858.CD009067

Rismawan, T., & Hartati, S. (2012). Case-Based Reasoning untuk Diagnosa Penyakit THT (Telinga Hidung dan Tenggorokan). *Indonesian Journal of Computing and Cybernetics Systems (IJCCS)*, *6*(2), 67–78. https://doi.org/10.22146/ijccs.2154

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of klaster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1177/003754977702900403

Tedjasukmana, P. 2012. Tata Laksana Hipertensi. *Cermin Dunia Kedokteran*, *39*(4), 251–255.

Triyanto, E. 2014. *Pelayanan Keperawatan bagi Penderita Hipertensi Secara Terpadu*. Yogyakarta: Graha Ilmu.

Whitworth, J. A. 2003. World Health Organization (WHO) International Society of Hypertension (ISH) statement on management of hypertension. *Journal of Hypertension*, *21*(11), 1983–1992. https://doi.org/10.1097/01.hjh.0000084751.37215.d2

Witten, I. H., & Frank, E. 2005. *Data mining: practical machine learning tools and techniques*. Burlington, USA: Morgan Kauffmann. https://doi.org/0120884070, 9780120884070

Zhu, S., Wu, J., Xiong, H., & Xia, G. 2011. Scaling up top-K cosine similarity search. *Data and Knowledge Engineering*, *70*(1), 60–83. https://doi.org/10.1016/j.datak.2010.08.004