ANALISIS RADIUS PADA ALGORITMA BIRCH BERDAMPAK TERHADAP DISTRIBUSI DAN KUALITAS CLUSTER

Yasir Hasan[1]

[1] *STMIK MULIA DARMA, Jl. H. Adam Malik, Padang Bulan, Rantau Utara, Labuhanbatu 21412*

ARTICLE INFORMATION

Received :February 2024
Revised: Juli 2024
Available online: Oktober, 2024

KEYWORDS

Birch, Cluster, Radius Parameters

CORRESPONDENCE

Phone: +62 81375557689
E-mail: yasirhasan.kom@gmail.com

A B S T R A C T

The BIRCH method is efficient in handling large data. However, the determination of the Radius (R) parameter, which is useful for determining the maximum radius of the cluster, must be considered. R values that are too small produce many small clusters (overclustering), while values that are too large produce clusters with high heterogeneity (underclustering). The R parameter affects the distribution results and the quality of the resulting clusters can be a major problem. The lack of clear guidance in determining the optimal R value can lead to the formation of inappropriate clusters or loss of important information in the data. This study aims to provide the impact of the R value on the distribution and quality of clusters in the BIRCH. Testing several R values on employee datasets that include non-linear distribution data. Analysis is carried out to identify the relationship between the R value and the resulting data distribution pattern. The evaluation results show that R values that are too small tend to produce over-clustering, while values that are too large cause under-clustering. The best cluster quality is achieved at a balanced R value, which is adjusted to the distribution and density of the employee dataset. Thus, it is important to choose the right R value to improve the performance of the BIRCH and ensure a representative cluster distribution. This finding provides practical guidance for adjusting the R parameter in clustering applications that implement the BIRCH.

PENDAHULUAN

Teknik *Clustering* memiliki peran penting dalam analisis data seperti seperti analisis pasar, deteksi anomali, bioinformatika, dan segmentasi gambar yang bertujuan pengelompokkan objek berdasarkan kesamaan karakteristik[1]. Algoritma *clustering* terus dikembangkan agar mampu menangani data yang besar dan kompleks, sehingga memberikan hasil yang lebih relevan dan bermanfaat dalam pengambilan keputusan berbasis data[2], [3], [4]. Seiring dengan semakin kompleksnya data yang dihasilkan, kebutuhan akan algoritma clustering yang efisien dan akurat juga semakin meningkat.

Algoritma *Balanced Iterative Reducing and Clustering using Hierarchies* (BIRCH) dikenal memiliki kemampuan dalam pengelompokkan data dalam skala besar dengan memori kecil dan waktu eksekusi yang cepat serta efisien[5], [6], [7]. BIRCH menggunakan pendekatan hierarkis yang menggabungkan kelebihan *clustering* berbasis partisi dan hierarki membentuk iterasi *Clustering Feature Tree* dalam mereduksi data dan memilki sumber daya komputasi yang terbatas. Namun, meskipun efisien, hasil *clustering* dengan BIRCH terkadang kurang akurat dipengaruhi oleh parameter radius maksimum *cluster* (R) dan jumlah *cluster* yang diinginkan, terutama pada dataset dengan distribusi non-linear atau *cluster* yang saling tumpang tindih[8], [9], [10]. Dalam algoritma BIRCH nilai R berfungsi sebagai batas maksimum jarak yang menentukan apakah suatu data dapat dimasukkan ke dalam *cluster* yang sudah ada atau harus membentuk *cluster* baru. Nilai *R* yang terlalu kecil dapat menghasilkan banyak *cluster* kecil, sementara nilai yang terlalu besar dapat menyebabkan *cluster* dengan heterogenitas tinggi[11], [12]. Menentukan nilai R yang tepat memerlukan pemahaman terhadap distribusi data, sifat dataset, serta tujuan analisis. Kebijaksanaan dalam menentukan R terletak pada pemahaman distribusi yaitu, dengan menghitung jarak maksimum (Dmaksimal) antar data ekstrem dan menggunakan persentase tertentu, seperti 50% dan 70% dari Dmaksimal, untuk mendapatkan keseimbangan antara jumlah *cluster* dan representasi data. Pendekatan ini perlu disertai validasi, seperti evaluasi dengan indeks kualitas *cluster*, untuk memastikan hasil yang sesuai dengan tujuan analisis[8], [13].

Hasil *cluster* yang kurang berkualitas dapat menyebabkan interpretasi data menjadi tidak optimal, yang berdampak pada keputusan atau tindakan yang diambil berdasarkan data tersebut[14]. Masalah ini muncul karena metode BIRCH lebih mengutamakan efisiensi dibandingkan ketepatan dalam menangkap struktur alami dari data[15]. Sebagai akibatnya, *cluster* yang dihasilkan dari atribut yang digunakan (Atr1, Atr2, Atr3, Atr4, Atr5) pola distribusi data tidak sepenuhnya linier, sehingga mempengaruhi akurasi hasil *clustering* dan sering kali tidak sepenuhnya merepresentasikan pola distribusi data yang sebenarnya. Selain itu, BIRCH juga rentan terhadap *noise* dan *outlier* yang dapat mengganggu pembentukan *cluster* sehingga hasil *clustering* yang dihasilkan terkadang kurang optimal.

Kualitas hasil *clustering* sangat penting ditentukan oleh pemilihan parameter dan metode yang digunakan. Kesalahan-kesalahan dalam menentukan jumlah cluster serta parameter lainnya mengakibatkan kegagalan dan ketidakoptimalan hasil *clustering*, seperti *over-clustering* terjadi ketika data dengan kemiripan besar berada dalam *cluster* yang berbeda, selain itu terdapat juga kondisi *under-clustering* data yang mana data yang berbeda atau dengan kemiripan kecil ditempatkan dalam *cluster* yang sama[8], [16]. Kedua kondisi ini tidak optimal dapat mengaburkan interpretasi data dan juga mengurangi efektivitas analisis. Sehingga, pemilihan nilai *R* yang tepat sangat penting untuk memastikan distribusi cluster yang representatif dan bermakna. Kualitas *cluster* yang dihasilkan oleh algoritma BIRCH memiliki dampak signifikan dalam sistem penempatan karyawan berdasarkan keahlian dan kinerja, penggunaan algoritma BIRCH dapat membantu mengelompokkan karyawan secara efisien, sehingga mempermudah proses penempatan dan meningkatkan produktivitas perusahaan. Namun, jika parameter *R* tidak dipilih dengan tepat, hasil *clustering* dapat menjadi kurang akurat, yang berpotensi mengarah pada keputusan yang tidak optimal.

Urgensi penelitian ini membahas pentingnya mengetahui pengaruh parameter R terhadap distribusi dan hasil kualitas *cluster* oleh algoritma BIRCH. Pemahaman akan dampak tersebut perlu diperhatikan praktisi data sehingga dapat melakukan penyesuaian

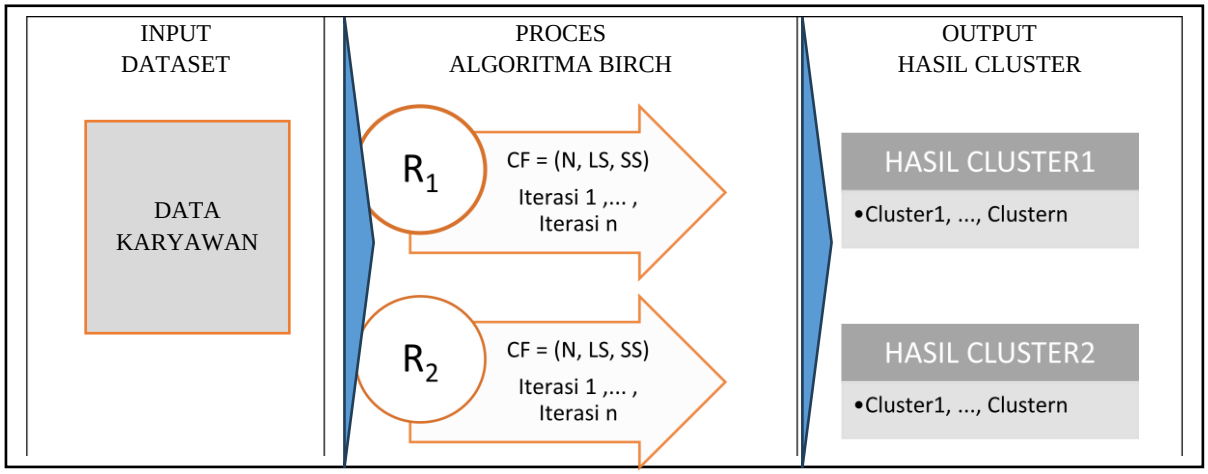
parameter yang tepat dalam peningkatan akurasi dan efisiensi proses *clustering* di berbagai aplikasi. Penelitian ini juga diharapkan dapat memberikan gambaran informatif untuk memilih parameter R yang sesuai dengan karakteristik data yang dianalisis.

METODE PENELITIAN

Penelitian ini berfokus pada evaluasi dampak parameter *R* dalam algoritma BIRCH terhadap distribusi dan kualitas *cluster*. Hasil uji dibandingkan dengan diilustrasikan melalui diagram.

2.1 Alur Penelitian

Pengujian analisis data ini dilakukan dengan memanfaatkan dataset berisi 30 data karyawan sebagai data input. Hasil akhir *cluster* terdiri dari tiga hasil akhir, yang mana untuk melakukan proses algoritma Birch menggunakan tiga nilai R yang berbeda. Nilai R yang berbeda ini dilakukan agar dapat mengevaluasi efektivitas pembentukan *cluster*. Nilai R pertama ditentukan sebesar 50% dari Dmaks awal, merepresentasikan tingkat toleransi yang lebih rendah sehingga menghasilkan *cluster* yang lebih tersegmentasi. Nilai R ketiga menggunakan 70% dari Dmaks awal, yang meningkatkan toleransi jarak antar data sehingga menghasilkan *cluster* yang lebih sedikit dan memiliki anggota yang banya. Hasil *cluster* dari ketiga nilai R tersebut dianalisis untuk melihat variasi jumlah dan kualitas *cluster* yang dihasilkan.



Gambar 1. Alur Penelitian

2.2 Analisis dan Validasi Hasil:

Data hasil eksperimen dianalisis untuk mengidentifikasi hubungan antara nilai *R*, distribusi *cluster*, dan kualitas *cluster* yang dihasilkan. Validasi dilakukan dengan membandingkan hasil *clustering* dari berbagai nilai *R*. Perbandingan ini akan diilustrasikan melalui visualisasi *cluster*.

HASIL DAN PEMBAHASAN

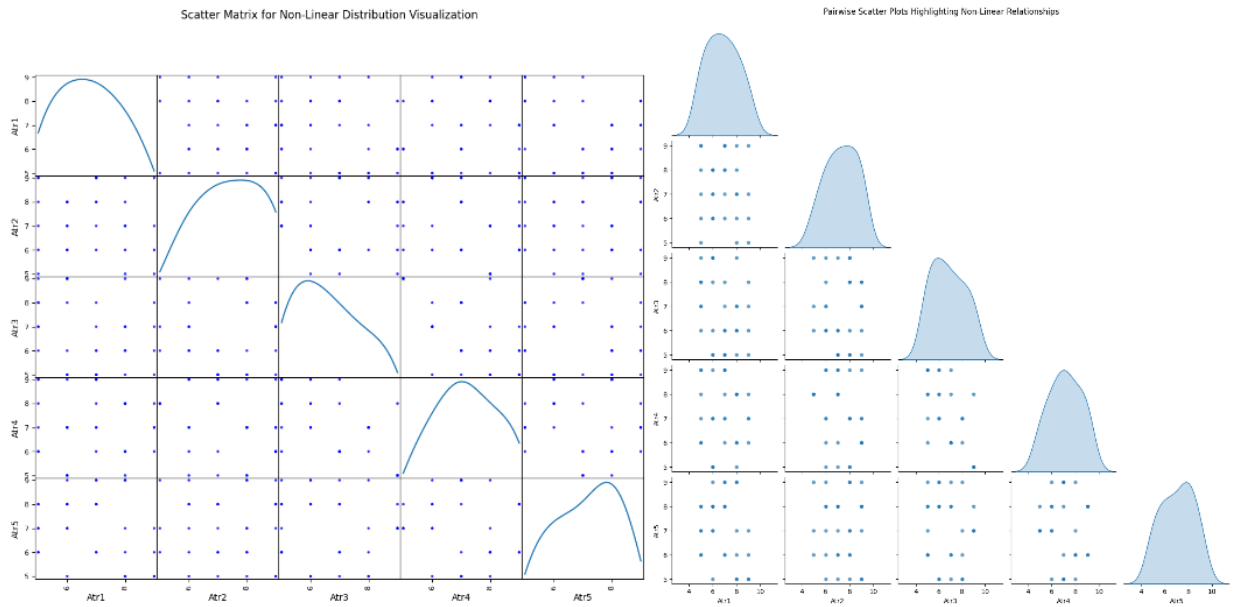
Dataset karyawan yang disajikan berisi 30 entri, di mana setiap entri mewakili satu karyawan dengan atribut-atribut yang menggambarkan karakteristik mereka. Dataset ini dapat merepresentasikan faktor seperti kinerja, kehadiran, keterampilan, produktivitas, atau nilai lain yang relevan dalam konteks analisis. Setiap karyawan memiliki nilai unik untuk kelima atribut ini. Misalnya, Karyawan1 memiliki nilai 6, 8, 5, 7, dan 9 untuk Atr1 hingga Atr5 secara berurutan. Dataset ini cocok untuk dianalisis menggunakan metode *clustering* seperti BIRCH untuk mengelompokkan karyawan berdasarkan kesamaan atribut mereka, yang dapat membantu dalam pengambilan keputusan manajemen seperti pembagian tim, evaluasi kinerja, atau strategi pengembangan karyawan.

Tabel 1. Data Karyawan

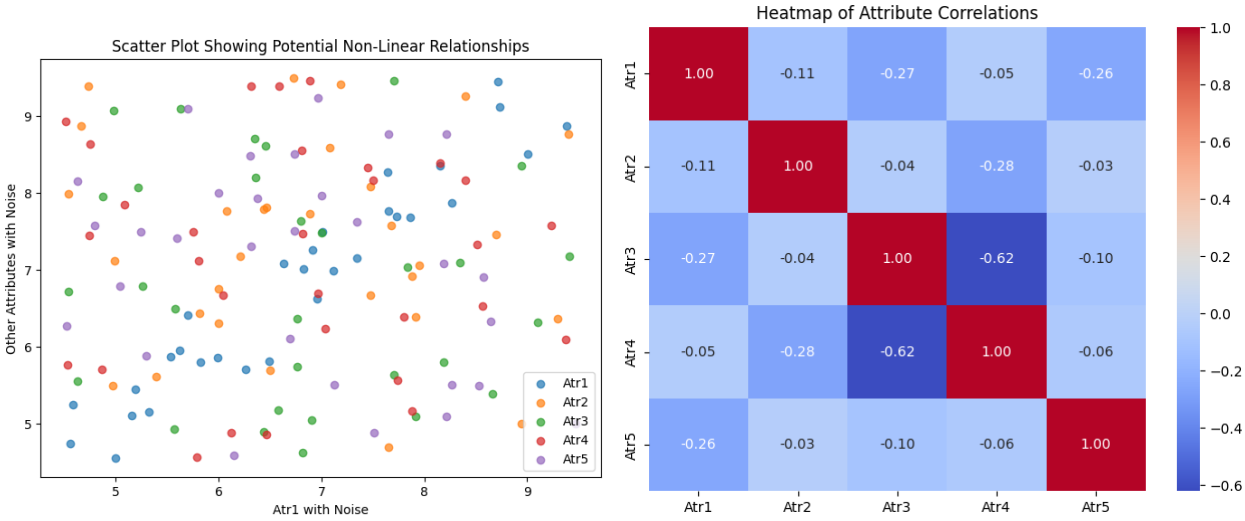
No.	Nama	Atr1	Atr2	Atr3	Atr4	Atr5
1	Karyawan1	6	8	5	7	9
2	Karyawan2	7	6	6	9	8
3	Karyawan3	9	5	7	8	6
4	Karyawan4	5	9	8	6	7
5	Karyawan5	6	6	9	5	7
6	Karyawan6	8	7	6	8	5
7	Karyawan7	7	9	7	6	8
8	Karyawan8	5	8	8	7	6

No.	Nama	Atr1	Atr2	Atr3	Atr4	Atr5
9	Karyawan9	8	5	6	8	9
10	Karyawan10	9	7	5	6	7
11	Karyawan11	6	7	9	5	8
12	Karyawan12	7	8	5	9	6
13	Karyawan13	9	9	6	7	5
14	Karyawan14	5	5	9	8	7
15	Karyawan15	8	6	7	6	9
16	Karyawan16	6	8	8	7	5
17	Karyawan17	7	9	5	9	8
18	Karyawan18	6	6	6	7	8
19	Karyawan19	7	7	5	8	6
20	Karyawan20	8	8	9	5	7
21	Karyawan21	5	9	7	6	8
22	Karyawan22	9	6	8	7	5
23	Karyawan23	6	7	5	9	8
24	Karyawan24	7	8	6	7	9
25	Karyawan25	5	6	7	9	6
26	Karyawan26	8	7	5	8	6
27	Karyawan27	7	9	8	7	9
28	Karyawan28	6	8	9	5	7
29	Karyawan29	5	7	6	9	8
30	Karyawan30	8	9	7	6	5

Dataset ini dapat termasuk dalam distribusi non-linear jika hubungan antara atribut (Atr1, Atr2, Atr3, Atr4, Atr5) tidak mengikuti pola linier yang sederhana. Dataset berbentuk tabel dengan baris mewakili karyawan dan kolom mewakili atribut (Data Tabular). Semua atribut (Atr1 hingga Atr5) berisi nilai numerik diskret dalam rentang 5-9, menunjukkan skala terurut tetapi tanpa deskripsi spesifik (Atribut Numerik). pola yang terbentuk dapat memperlihatkan cluster atau kelompok yang tidak dapat dipisahkan secara linier seperti hubungan berbentuk melingkar atau acak, dapat dilihat karyawan dengan nilai Atr1 tinggi tidak selalu memiliki nilai Atr2 tinggi, ini menunjukkan kemungkinan distribusi non-linear walau dalam rentang 5-9 (Distribusi Non-Linear). Terdiri dari banyak variabel (lima atribut), sehingga analisis hubungan antar atribut menjadi kompleks (Dataset Multivariat).



Gambar 2. Diagram Scatter Plot, Pairplot, Line Plot dan Heatmap mengidentifikasi distribusi atribut



Gambar 3. Diagram Scatter Plot, Pairplot, Line Plot dan Heatmap mengidentifikasi distribusi atribut

Proses Algoritma BIRCH

BIRCH untuk *clustering* menggunakan struktur hierarki yang mana secara matematisnya, elemen utama dalam algoritma ini meliputi pembentukan CF *Tree* (*Clustering Feature Tree*)[8], dengan setiap node menyimpan informasi *Clustering Feature* (CF).

1. Inisialisasi *Cluster*

   Setiap node dalam CF *Tree* dirangkum oleh tiga elemen utama: CF = (N, LS, SS), berikut pendefenisian untuk data Karyawan1 sebagai *cluster* awal**:**

   CF1 = (N, LS, SS) → N=1, LS = [6,8,5,7,9], SS = [6^2,8^2,5^2,7^2,9^2] = [36,64,25,49,81]

2. Tambahkan Data Baru

   Untuk setiap data berikutnya, hitung jarak antar data menggunakan jarak Euclidean[10]:

   $$D = \sqrt{\sum_{i=1}^{d}(xi - yi)^2} \quad (1)$$

   Karyawan1(x): [6,8,5,7,9] ke Karyawan2(y): [5,5,9,8,7]

   $d(x,y)=\sqrt{(x1-y1)^2+(x2-y2)^2+\cdots+(xn-yn)^2}$

   Jarak Karyawan1 ke Karyawan2 = $\sqrt{(6-5)^2+(8-5)^2+(5-9)^2+(7-8)^2+(9-7)^2} = 33.2$

   Selanjutnya, jarak karyawan1 ke setiap karyawan secara keseluruhan dapat dilihat pada tabel 2.

   Tabel 2. Jarak antar data

Nama	Distance	Nama	Distance	Nama	Distance
Karyawan1	1.66	Karyawan11	4.66	Karyawan21	3.57
Karyawan2	1.66	Karyawan12	2.96	Karyawan22	5.17
Karyawan3	4.33	Karyawan13	4.87	Karyawan23	1.32
Karyawan4	4.33	Karyawan14	4.56	Karyawan24	1.66
Karyawan5	4.97	Karyawan15	3.12	Karyawan25	3.57
Karyawan6	3.84	Karyawan16	4.56	Karyawan26	2.96
Karyawan7	3.28	Karyawan17	2.40	Karyawan27	3.43
Karyawan8	4.09	Karyawan18	1.66	Karyawan28	4.97
Karyawan9	2.60	Karyawan19	2.60	Karyawan29	1.94
Karyawan10	3.57	Karyawan20	5.17	Karyawan30	4.97

   Jika jarak antara data baru dan *centroid cluster* yang ada lebih kecil dari radius (R), tambahkan data tersebut ke *cluster*. Jika tidak, buat *cluster* baru.

3. Inisialisasi *Threshold* (*Centroid*) dan Nilai Radius (R) pada BIRCH

   Threshold

   Threshold untuk menentukan kapan dua data atau cluster dapat digabungkan[9]. Berikut gunakan radius atau jarak Euclidean antara *centroid cluster*. Menghitung *centroid* dari sebuah *cluster* karyawan1 dan karyawan2:

   $$Cawal = \frac{LSbaru(LSkaryawan1 + LSkaryawan2)}{Nbaru(Nkaryawan1 + Nkaryawan2)}$$

   $$Cawal = \frac{(6,8,5,7,9)+(7,6,6,9,8)}{1+1} = \frac{(13,14,11,16,17)}{2} = (6.5, 7, 5.5, 8, 8.5)$$

Radius mengukur seberapa jauh data dalam *cluster* tersebar dari *centroid*. Nilai R adalah parameter utama dalam algoritma BIRCH yang menentukan jarak maksimum di mana sebuah data baru dapat ditambahkan ke *cluster* yang ada. Besarnya R biasanya ditentukan berdasarkan karakteristik dataset, seperti distribusi data, dimensi, dan variasi antar data.
Secara umum R didapatkan berdasarkan rumus:

$$Radius = \sqrt{\frac{SS}{SN} - \left(\frac{SS^2}{SN}\right)} \quad (2)$$

Namun, penggunaan rumus ini akan menyebabkan jumlah *cluster* melebar atau banyak dan *cluster* anggota sedikit. Untuk itu pendefenisian R yang digunakan dipilih dengan dua penentu seperti disebutkan sebelumnya di awal. Pada tabel 2 diketahui nilai tertinggi yang akan digunakan sebagai Dmaks=5.17 yang terdapat pada karyawan20 dan karyawan22

Pendefenisian R(1) untuk perhitungan BIRCH pertama

Menggunakan skala jarak 50% = 0.5 dari karakteristik semua jarak antar data.:
$R(1) = 0.5 * Dmaks = 0.5 * 5.17 = 2.59$

Pendefenisian R(2) untuk perhitungan BIRCH kedua

Menggunakan skala jarak 70% = 0.7 dari karakteristik semua jarak antar data.:
$R(2) = 0.5 * Dmaks = 0.5 * 5.17 = 3.62$

4. Pembentukan *Cluster*

Pembentukan *Cluster* menggunakan R(1)
Pada langkah berikut ini mendapatkan jarak setiap baris data karyawan dengan *centroid* awal.
$Centroid\ Awal \rightarrow$ Cawal $= (6.5, 7, 5.5, 8, 8.5)$
Centroid Awal ke Karyawan 1

D(Karyawan 1, Cawal)= $\sqrt{(6-6.5)^2 + (8-7)^2 + (5-5.5)^2 + (7-8)^2 + (9-8.5)^2} = 1.66$

Centroid Awal ke Karyawan 2

D(Karyawan 2, Cawal)= $\sqrt{(7-6.5)^2 + (6-7)^2 + (6-5.5)^2 + (9-8)^2 + (8-8.5)^2} = 1.66$

Centroid Awal ke Karyawan 3

D(Karyawan 3, Cawal)= $\sqrt{(9-6.5)^2 + (5-7)^2 + (7-5.5)^2 + (8-8)^2 + (6-8.5)^2} = 4.33$

Selanjutnya pada tabel di bawah dapat dilihat LS' yaitu jarak *centroid* awal ke setiap nilai karyawan.

Tabel 3. Jarak *centroid* awal ke setiap nilai karyawan

Nama	LS'					D
karyawan1	0.25	1.00	0.25	1.00	0.25	1.66
Karyawan2	0.25	1.00	0.25	1.00	0.25	1.66
Karyawan3	6.25	4.00	2.25	0.00	6.25	4.33
Karyawan4	2.25	4.00	6.25	4.00	2.25	4.33
Karyawan5	0.25	1.00	12.25	9.00	2.25	4.97
Karyawan6	2.25	0.00	0.25	0.00	12.25	3.84
Karyawan7	0.25	4.00	2.25	4.00	0.25	3.28
Karyawan8	2.25	1.00	6.25	1.00	6.25	4.09
Karyawan9	2.25	4.00	0.25	0.00	0.25	2.60
Karyawan10	6.25	0.00	0.25	4.00	2.25	3.57
Karyawan11	0.25	0.00	12.25	9.00	0.25	4.66
Karyawan12	0.25	1.00	0.25	1.00	6.25	2.96
Karyawan13	6.25	4.00	0.25	1.00	12.25	4.87
Karyawan14	2.25	4.00	12.25	0.00	2.25	4.56
Karyawan15	2.25	1.00	2.25	4.00	0.25	3.12
Karyawan16	0.25	1.00	6.25	1.00	12.25	4.56
Karyawan17	0.25	4.00	0.25	1.00	0.25	2.40
Karyawan18	0.25	1.00	0.25	1.00	0.25	1.66
Karyawan19	0.25	0.00	0.25	0.00	6.25	2.60
Karyawan20	2.25	1.00	12.25	9.00	2.25	5.17
Karyawan21	2.25	4.00	2.25	4.00	0.25	3.57
Karyawan22	6.25	1.00	6.25	1.00	12.25	5.17

Nama	LS’					D
Karyawan23	0.25	0.00	0.25	1.00	0.25	1.32
Karyawan24	0.25	1.00	0.25	1.00	0.25	1.66
Karyawan25	2.25	1.00	2.25	1.00	6.25	3.57
Karyawan26	2.25	0.00	0.25	0.00	6.25	2.96
Karyawan27	0.25	4.00	6.25	1.00	0.25	3.43
Karyawan28	0.25	1.00	12.25	9.00	2.25	4.97
Karyawan29	2.25	0.00	0.25	1.00	0.25	1.94
Karyawan30	2.25	4.00	2.25	4.00	12.25	4.97

Klaster ditentukan berdasarkan jarak (D) pada tabel yang di atas, dengan menggunakan nilai radius R(1) = 2.59 yang sudah didapatkan sebelumnya. Penentuan klaster dengan cara D <= R(1) adalah anggota klaster.

Tabel 4. Penentuan klaster

Nama	D	Klaster	Nama	D	Klaster
karyawan1	1.66	cluster1	Karyawan16	4.56	non
Karyawan2	1.66	cluster1	Karyawan17	2.40	cluster1
Karyawan3	4.33	non	Karyawan18	1.66	cluster1
Karyawan4	4.33	non	Karyawan19	2.60	non
Karyawan5	4.97	non	Karyawan20	5.17	non
Karyawan6	3.84	non	Karyawan21	3.57	non
Karyawan7	3.28	non	Karyawan22	5.17	non
Karyawan8	4.09	non	Karyawan23	1.32	cluster1
Karyawan9	2.60	non	Karyawan24	1.66	cluster1
Karyawan10	3.57	non	Karyawan25	3.57	non
Karyawan11	4.66	non	Karyawan26	2.96	non
Karyawan12	2.96	non	Karyawan27	3.43	non
Karyawan13	4.87	non	Karyawan28	4.97	non
Karyawan14	4.56	non	Karyawan29	1.94	cluster1
Karyawan15	3.12	non	Karyawan30	4.97	non

Dari tabel di atas diketahui karyawan-karyawan yang memiliki nilai lebih kecil sama dengan R(1) adalah *cluster*1.
*Cluster*1 : Karyawan1, Karyawan2, Karayawan17, Karyawan18, Karyawan23, Karyawan24

Evaluasi Data Sisa:
Kumpulkan data yang bukan anggota *Cluster* 1 dan bentuk pusat *cluster* baru dengan memilih data untuk klaster baru. Pada tahap berikut dipilih data selanjutnya, yaitu data Karyawan3 dan Karyawan4 untuk membuat *centroid* baru. Melakukan cara yang sama dengan tahapan-tahapan sebelumnya, dimulai dari Hitung jarak pusat baru, Penentuan anggota klaster baru, dan pembaruan CF. Dari penerapan R(1) didapat 10 klaster dengan 9 iterasi. Berikut hasil 10 klaster tersebut:

Tabel 5. *Cluster* yang terbentuk dari R(1)

No.	Nama Klaster	Anggota Klaster
1	*Cluster 1*	Karyawan1, Karyawan2, Karyawan18, Karyawan23, Karyawan24, Karyawan29
2	*Cluster 2*	Karyawan3, Karyawan22
3	*Cluster 3*	Karyawan4, Karyawan7, Karyawan8, Karyawan11, Karyawan16, Karyawan21, Karyawan27, Karyawan28
4	*Cluster 4*	Karyawan5, Karyawan20
5	*Cluster 5*	Karyawan6, Karyawan12, Karyawan13, Karyawan19, Karyawan26
6	*Cluster 6*	Karyawan9, Karyawan15
7	*Cluster 7*	Karyawan10
8	*Cluster 8*	Karyawan14, Karyawan25
9	*Cluster 9*	Karyawan17
10	*Cluster 10*	Karyawan30

Pembentukan *Cluster* menggunakan R(2)

Tahapan pembuatan centroid awal bagian ini sama dengan tahapan sebelumnya, begitupun untuk mendapatkan D setiap karyawan. Berikut ini mendapatkan jarak setiap baris data karyawan dengan *centroid* awal.

$Centroid\ Awal \rightarrow$ Cawal $= (6.5, 7, 5.5, 8, 8.5)$

Centroid Awal ke Karyawan 1

D(Karyawan 1, Cawal)= $\sqrt{(6-6.5)^2 + (8-7)^2 + (5-5.5)^2 + (7-8)^2 + (9-8.5)^2} = 1.66$

Centroid Awal ke Karyawan 2

D(Karyawan 2, Cawal)= $\sqrt{(7-6.5)^2 + (6-7)^2 + (6-5.5)^2 + (9-8)^2 + (8-8.5)^2} = 1.66$

Centroid Awal ke Karyawan 3

D(Karyawan 3, Cawal)= $\sqrt{(9-6.5)^2 + (5-7)^2 + (7-5.5)^2 + (8-8)^2 + (6-8.5)^2} = 4.33$

Dikarenakan sama tahapan pembentukkan D, maka langsung dilakukan perbandingan D dengan R(2): R(2) = Dmaks*70%. = 5.17*70% = 3.62. Lakukan pembandingan nilai D setiap karyawan dengan R(2), Berikut hasil perbandingan

Tabel 6. Jarak centroid awal ke setiap nilai karyawan

Nama	LS'					D	Klaster
Karyawan1	0.25	1.00	0.25	1.00	0.25	1.66	cluster1
Karyawan2	0.25	1.00	0.25	1.00	0.25	1.66	cluster1
Karyawan3	6.25	4.00	2.25	0.00	6.25	4.33	non
Karyawan4	2.25	4.00	6.25	4.00	2.25	4.33	non
Karyawan5	0.25	1.00	12.25	9.00	2.25	4.97	non
Karyawan6	2.25	0.00	0.25	0.00	12.25	3.84	non
Karyawan7	0.25	4.00	2.25	4.00	0.25	3.28	cluster1
Karyawan8	2.25	1.00	6.25	1.00	6.25	4.09	non
Karyawan9	2.25	4.00	0.25	0.00	0.25	2.60	cluster1
Karyawan10	6.25	0.00	0.25	4.00	2.25	3.57	cluster1
Karyawan11	0.25	0.00	12.25	9.00	0.25	4.66	non
Karyawan12	0.25	1.00	0.25	1.00	6.25	2.96	cluster1
Karyawan13	6.25	4.00	0.25	1.00	12.25	4.87	non
Karyawan14	2.25	4.00	12.25	0.00	2.25	4.56	non
Karyawan15	2.25	1.00	2.25	4.00	0.25	3.12	cluster1
Karyawan16	0.25	1.00	6.25	1.00	12.25	4.56	non
Karyawan17	0.25	4.00	0.25	1.00	0.25	2.40	cluster1
Karyawan18	0.25	1.00	0.25	1.00	0.25	1.66	cluster1
Karyawan19	0.25	0.00	0.25	0.00	6.25	2.60	cluster1
Karyawan20	2.25	1.00	12.25	9.00	2.25	5.17	non
Karyawan21	2.25	4.00	2.25	4.00	0.25	3.57	cluster1
Karyawan22	6.25	1.00	6.25	1.00	12.25	5.17	non
Karyawan23	0.25	0.00	0.25	1.00	0.25	1.32	cluster1
Karyawan24	0.25	1.00	0.25	1.00	0.25	1.66	cluster1
Karyawan25	2.25	1.00	2.25	1.00	6.25	3.57	cluster1
Karyawan26	2.25	0.00	0.25	0.00	6.25	2.96	cluster1
Karyawan27	0.25	4.00	6.25	1.00	0.25	3.43	cluster1
Karyawan28	0.25	1.00	12.25	9.00	2.25	4.97	non
Karyawan29	2.25	0.00	0.25	1.00	0.25	1.94	cluster1
Karyawan30	2.25	4.00	2.25	4.00	12.25	4.97	non

Dari tabel di atas diketahui karyawan-karyawan yang memiliki nilai lebih kecil sama dengan R(2) adalah *cluster1*.

Cluster1: Karyawan1, Karyawan2, Karyawan7, Karyawan9, Karyawan10, Karyawan12, Karyawan15, Karyawan17, Karyawan18, Karyawan19, Karyawan21, Karyawan23, Karyawan24, Karyawan25, Karyawan26, Karyawan27, Karyawan29

Evaluasi Data Sisa:

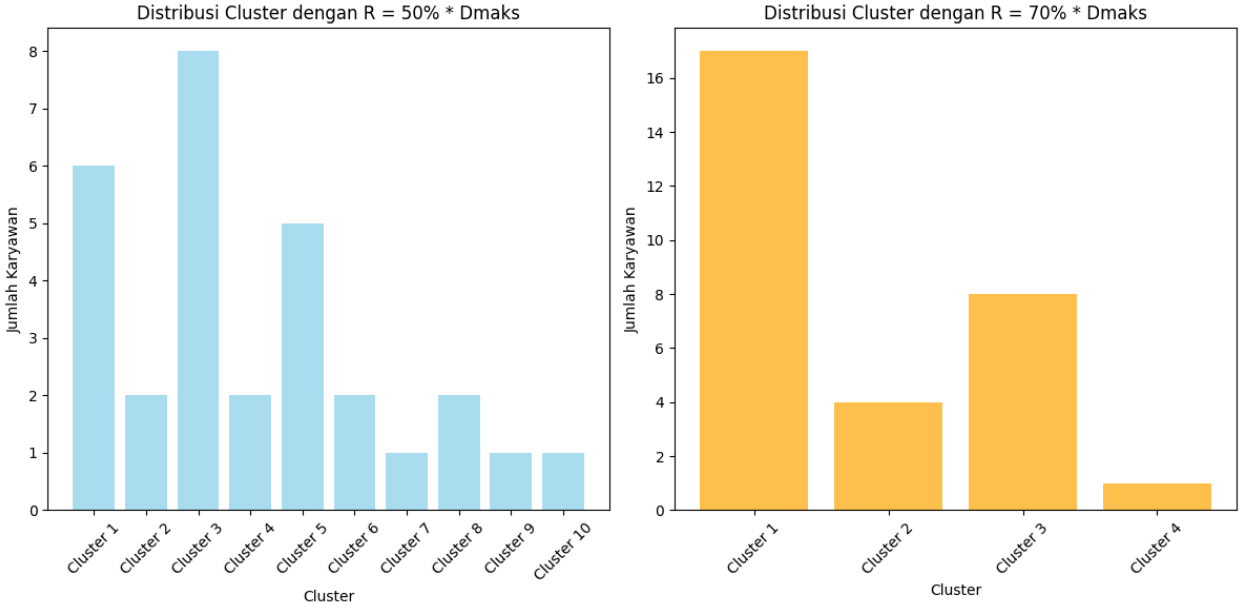
Seperti sebelumnya cara yang sama dilakukan untuk mengumpulkan data yang bukan anggota *Cluster* 1 dan bentuk pusat *cluster* baru dengan memilih data untuk klaster baru. Pada tahap ini dipilih data selanjutnya, yaitu data Karyawan3 dan Karyawan4 untuk membuat *centroid* baru. Melakukan cara yang sama dengan tahapan-tahapan sebelumnya, dimulai dari Hitung jarak pusat baru, Penentuan anggota klaster baru, dan pembaruan CF. Dari penerapan R(2) didapat 4 klaster dengan 3 iterasi.

Tabel 7. Cluster yang terbentuk dari R(2)

No.	Nama Klaster	Anggota Klaster
1	Cluster 1	Karyawan1, Karyawan2, Karyawan7, Karyawan9, Karyawan10, Karyawan12, Karyawan15, Karyawan17, Karyawan18, Karyawan19, Karyawan21, Karyawan23, Karyawan24, Karyawan25, Karyawan26, Karyawan27, Karyawan29
2	Cluster 2	Karyawan3, Karyawan6, Karyawan13, Karyawan22
3	Cluster 3	Karyawan4, Karyawan5, Karyawan8, Karyawan11, Karyawan16, Karyawan20, Karyawan28, Karyawan30
4	Cluster 4	Karyawan14

5. Evaluasi

Pada bab ini, dilakukan evaluasi terhadap hasil penerapan algoritma BIRCH dengan dua nilai parameter radius (R) yang berbeda, yaitu R1 = 50% dari Dmaks dan R2 = 70% dari *Dm*aks. Pemilihan nilai R ini bertujuan untuk mengetahui pengaruh besar radius terhadap jumlah *cluster* yang dihasilkan serta kualitas distribusi data di dalam *cluster*.



Grafik 1. Hasil Cluster

Hasil *Clustering* dengan R1 = 50%*Dmaks menghasilkan 10 *cluster*. Berikut adalah distribusi karyawan pada masing-masing *cluster*:
Cluster 1: Karyawan1, Karyawan2, Karyawan18, Karyawan23, Karyawan24, Karyawan29
Cluster 2: Karyawan3, Karyawan22
Cluster 3: Karyawan4, Karyawan7, Karyawan8, Karyawan11, Karyawan16, Karyawan21, Karyawan27, Karyawan28
Cluster 4: Karyawan5, Karyawan20
Cluster 5: Karyawan6, Karyawan12, Karyawan13, Karyawan19, Karyawan26
Cluster 6: Karyawan9, Karyawan15
Cluster 7: Karyawan10
Cluster 8: Karyawan14, Karyawan25
Cluster 9: Karyawan17
Cluster 10: Karyawan30
Analisis awal menunjukkan bahwa nilai R yang kecil (R1) menghasilkan jumlah *cluster* yang lebih banyak. Hal ini mengindikasikan bahwa data lebih tersegmentasi.

Hasil *Clustering* dengan R2 = 70%
Hasil *clustering* dengan R=70%·Dmaks menghasilkan 4 *cluster*. Berikut adalah distribusi karyawan pada masing-masing *cluster*:
Cluster 1: Karyawan1, Karyawan2, Karyawan7, Karyawan9, Karyawan10, Karyawan12, Karyawan15, Karyawan17, Karyawan18, Karyawan19, Karyawan21, Karyawan23, Karyawan24, Karyawan25, Karyawan26, Karyawan27, Karyawan29
Cluster 2: Karyawan3, Karyawan6, Karyawan13, Karyawan22
Cluster 3: Karyawan4, Karyawan5, Karyawan8, Karyawan11, Karyawan16, Karyawan20, Karyawan28, Karyawan30
Cluster 4: Karyawan14
Dengan nilai R yang lebih besar (R2), jumlah *cluster* berkurang, menunjukkan pengelompokan data yang lebih luas.

KESIMPULAN

Evaluasi ini difokuskan untuk pembahasan jumlah *Cluster* dan Distribusi Data dalam *Cluster*, berikut kesimpulan yang dihasilkan

1. Ketika menggunakan R1 = 50%, data menghasilkan lebih banyak *cluster* (10 cluster). Sebaliknya, dengan R2 = 70%, data terkelompok menjadi lebih sedikit (4 *cluster*). Hal ini mengindikasikan bahwa parameter R secara langsung memengaruhi granularitas pengelompokan data. Nilai R yang kecil (R1) menghasilkan jumlah *cluster* yang lebih banyak dan granular, sementara nilai R yang besar (R2) cenderung menghasilkan jumlah *cluster* yang lebih sedikit tetapi lebih luas.
2. Pada R(1), distribusi data lebih spesifik dan berfokus pada keseragaman kecil dalam jarak antar data. Namun, pada R(2), beberapa *cluster* terlihat memiliki anggota yang lebih banyak, menunjukkan pengelompokan data yang lebih luas. Pemilihan nilai R yang tepat bergantung pada kebutuhan analisis, seperti seberapa detail segmentasi data yang diinginkan.

REFERENCES

[1] H. Sunandar, "Machine Learning Pengenalan Anura Berdasarkan Corak dan Warna," 2023. doi: 10.54367.

[2] J. Yang, Z. Sun, and Y. Chen, "Fault detection using the clustering-kNN rule for gas sensor arrays," *Sensors (Switzerland)*, vol. 16, no. 12, Dec. 2016, doi: 10.3390/s16122069.

[3] F. E. Ozturk, N. Demirel, and M. Bilgisi, "Comparison of the Methods to Determine Optimal Number of Cluster," 2023. [Online]. Available: www.dergipark.gov.tr/veri

[4] F. Ghifari, H. Rachmat, D. Sukma, and E. Atmaja, "DESIGN OF AUTOMATION INSPECTION SYSTEM USING CLUSTER IDENTIFICATION METHOD BASED ON LEATHER SHOES COLOUR AT VENAMON CORPORATION," 2015.

[5] J. Lei, "An extended BIRCH-based clustering algorithm for large time-series datasets," Aug. 2016.

[6] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," 1996.

[7] B. Lorbeer, A. Kosareva, B. Deva, D. Softić, P. Ruppel, and A. Küpper, "Variations on the Clustering Algorithm BIRCH," *Big Data Research*, vol. 11, pp. 44–53, Mar. 2018, doi: 10.1016/j.bdr.2017.09.002.

[8] D. Kathiravan, "Sentence-Similarity Based Document Clustering Using Birch Algorithm," *International Journal of Innovative Research in Computer and Communication Engineering (An ISO*, vol. 3297, no. 5, 2007, doi: 10.15680/ijircce.2015.0305055.

[9] B. Yu and J. Xiong, "A Novel WSN Traffic Anomaly Detection Scheme Based on BIRCH," *Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology*, vol. 44, no. 1, pp. 305–313, Jan. 2022, doi: 10.11999/JEIT201004.

[10] A. Alzu'Bi and M. Barham, "Automatic BIRCH thresholding with features transformation for hierarchical breast cancer clustering," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1498–1507, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1498-1507.

[11] A. R. Rizalde, H. A. Mubarak, G. Ramadhan, and Mohd. A. Fatan, "Comparison of K-Means, BIRCH and Hierarchical Clustering Algorithms in Clustering OCD Symptom Data," *Public Research Journal of Engineering, Data Technology and Computer Science*, vol. 1, no. 2, pp. 102–108, Feb. 2024, doi: 10.57152/predatecs.v1i2.1106.

[12] M. C. Nwadiugwu, "Gene-Based Clustering Algorithms: Comparison Between Denclue, Fuzzy-C, and BIRCH," 2020, *SAGE Publications Inc.* doi: 10.1177/1177932220909851.

[13] X. Xia, "Clustering Analysis of Interactive Learning Activities Based on Improved BIRCH Algorithm."

[14] Z. Yan, G. Yang, R. He, H. Yang, H. Ci, and R. Wang, "Ship Trajectory Clustering Based on Trajectory Resampling and Enhanced BIRCH Algorithm," *J Mar Sci Eng*, vol. 11, no. 2, Feb. 2023, doi: 10.3390/jmse11020407.

[15] M. Mohammad, B. Supervisor, A. Gazi, and A. ' Bi, "An Improved BIRCH Algorithm for Breast Cancer Clustering تصنيف مرض سرطان الثدي باستخدام خوارزمية (BIRCH(المحسنة 2020" .Prepared By,

[16] M. Malarczyk, S. Katsura, M. Kaminski, and K. Szabat, "A Novel Meta-Heuristic Algorithm Based on Birch Succession in the Optimization of an Electric Drive with a Flexible Shaft," *Energies (Basel)*, vol. 17, no. 16, Aug. 2024, doi: 10.3390/en17164104.