
ANALISIS CLUSTER HASIL TRY OUT SISWA MTS ALHUDA GORONTALO DENGAN CHI-SIM COSIMILARITY DAN K-MEANS CLUSTERING

(*CLUSTER ANALYSIS OF RESULT TRY OUT OF STUDENTS MTS
ALHUDA GORONTALO WITH CHI-SIM COSIMILARITY AND
K-MEANS CLUSTERING*)

Fahrezal Zubedi

Universitas Negeri Gorontalo, fahrezal@ung.ac.id

Abstrak

Tujuan penelitian ini yaitu menemukan kelompok siswa dan kelompok mata pelajaran yang homogen sehingga bisa memantau atau mengetahui kinerja akademik siswa. Langkah pertama yaitu mentransformasi data dengan menggunakan normalisasi min-max. Setelah itu, diterapkan χ -Sim co-similarity untuk menghasilkan matriks similaritas siswa (SS) dan similaritas pelajaran (SP). Masing-masing SS dan SP dikelompokkan menggunakan algoritma *k-means clustering* dan menggunakan *Silhouette* untuk menentukan banyaknya kelompok yang terbaik. Pada pengelompokan SS diperoleh nilai *Silhouette* terbesar yaitu 0,9755781 pada iterasi keempat yang mempartisi menjadi 4 *cluster* sebagai berikut 67 siswa pada *cluster* 1, 9 siswa pada *cluster* 2, 45 siswa pada *cluster* 3 dan 43 siswa pada *cluster* 4. Pada SP diperoleh nilai *Silhouette* terbesar yaitu 0,5756133 pada iterasi keempat yang mempartisi menjadi 2 *cluster* sebagai berikut Bahasa Indonesia dan Bahasa Inggris pada *cluster* 1 dan Matematika dan IPA pada *cluster* 2.

Kata kunci: *Clustering, χ -sim co-similarity, K-means.*

Abstract

The purpose of this paper is to find clusters of students and clusters of subjects that are homogeneous so that we can monitor or find out students' academic performance. The first step is to transform data using min-max normalization. After that, χ -Sim co-similarity is applied to produce a similarity matrix of student (SS) and similarity matrix lesson (SP). Each SS and SP are grouped using the k-means clustering algorithm and using Silhouette to determine the best number of clusters. In the similarity matrix students the largest Silhouette value obtained was 0.9755781 in the fourth iteration which partitioned 4 clusters as follows 67 students in cluster 1, 9 students in cluster 2, 45 students in cluster 3 and 43 students in cluster 4. In the similarity matrix lessons the largest Silhouette value obtained was 0,5756133 in the fourth iteration which partitioned 2 clusters as follows Indonesian and English in cluster 1 and Mathematics and science in cluster 2.

Keywords: *Clustering, χ -sim co-similarity, K-means*

PENDAHULUAN

Nilai rata-rata merupakan indikator yang digunakan untuk mengevaluasi perkembangan kinerja akademik siswa. Banyak sekolah menetapkan minimum nilai rata-rata yang harus dipertahankan untuk melanjutkan ke jenjang selanjutnya. Namun, penggunaan nilai rata-rata yang dijadikan sebagai tolak ukur evaluasi perkembangan kinerja akademik merupakan metode konvensional. Pengajar kesulitan dalam mengelompokkan siswa yang berjumlah banyak ke beberapa kelas/kelompok dengan kriteria tertentu dengan menggunakan metode konvensional. Hal ini membutuhkan waktu penyelesaian yang sangat lama. Pengelompokan secara manual memberikan hasil variasi nilai yang tidak merata di tiap-tiap kelompok.

Ada beberapa teknik dalam statistika yang dapat mengevaluasi perkembangan kinerja akademik dari setiap siswa. Salah satu teknik tersebut adalah analisis *cluster*. Analisis *cluster* dapat digunakan untuk mengidentifikasi kelompok siswa atau mata pelajaran yang homogen satu sama lain berdasarkan pengukuran kesamaan. *Clustering* adalah mengelompokkan objek menjadi beberapa *cluster*, sehingga objek yang memiliki kemiripan tinggi terhadap objek lain dikelompokkan ke dalam satu *cluster*, sedangkan objek yang tidak terdapat kemiripan terhadap objek lain dikelompokkan ke *cluster* yang lainnya (Han, 2011). *Clustering* adalah proses klasifikasi data yang mempartisi data sesuai jumlah kriteria untuk mengetahui pola-pola yang ada pada *dataset* (Liu, Xin, Cheung, & Yan, 2014). Pengelompokan objek-objek ke dalam kelompok bertujuan menemukan pengelompokan yang optimal, yaitu *clusters* yang terbentuk memiliki homogenitas *internal* dan heterogenitas *eksternal* yang tinggi (Bergeron, 2002). Terdapat beberapa algoritma *clustering* untuk mengelompokkan data. Salah satunya yaitu Algoritma *K-means clustering*.

K-Means clustering merupakan algoritma *clustering* yang mengelompokkan data berdasarkan kedekatan antar objek sesuai dengan jarak *Euclidean*. Kelebihan dari algoritma *k-means* yaitu mudah untuk diimplementasikan, waktu yang dibutuhkan untuk menjalankan algoritma ini relatif cepat, mudah diadaptasi, dan umum digunakan (X. Wu and VKumar, 2009). Selain itu, teknik *clustering* populer lainnya yang terbukti mampu dalam mengelompokkan data, diantaranya yaitu *K-Means*, *Partitioning Around Medoids* (PAM), dan *Fuzzy C-Means*. Untuk penggunaan *euclidean distance*, *k-means* memberikan akurasi yang lebih baik daripada PAM dan *Fuzzy C-Means* (Mondal & Paul Choudhury, 2013). Hussain telah membahas sebuah ukuran baru, yang disebut χ -*Sim cosimilarity* pada penerapan *data mining* dan bioinformatika untuk menghasilkan matriks similaritas gen atau dokumen (baris) dan similaritas kondisi atau kata (kolom) sebelum dilakukan proses *clustering* (Hussain & Ramazan, 2016). Alhadi juga telah melakukan penelitian terkait χ -*Sim cosimilarity* yang dikombinasikan dengan *Agglomerative hierarchical clustering* yang diterapkan pada data ekspresi gen (Bustamam, Zubedi, & Siswantining, 2018). Oleh karena itu, penelitian ini mengusulkan χ -*Sim cosimilarity* dan algoritma *k-means clustering* pada data hasil *Try Out* Siswa yang bertujuan untuk menemukan kelompok siswa dan kelompok mata pelajaran yang homogen sehingga bisa memantau atau mengetahui kinerja akademik siswa.

Normalisasi Min-Max

Normalisasi *min-max* mempertahankan hubungan antar nilai pada data asli dan mengubah entri-entri matriks ke interval $[0,1]$ (Chin, Mirzal, & Harona, 2015). Berikut merupakan rumus normalisasi min-max (Patro & sahu, 2015):

$$x'_{(a,b)} = \frac{x_{(a,b)} - \min(x_a)}{\max(x_a) - \min(x_a)} * (new_{max} - new_{min}) + new_{min} \quad (1)$$

Keterangan:

- $x'_{(a,b)}$: nilai elemen (a, b) pada matriks data setelah dinormalisasi
- $x_{(a,b)}$: nilai elemen (a, b) pada matriks data sebelum dinormalisasi
- $\min(x_a)$: nilai minimum pada baris ke- a matriks data sebelum dinormalisasi
- $\max(x_a)$: nilai maksimum pada baris ke- a matriks data sebelum dinormalisasi
- new_{max} : batas nilai maksimum baru
- new_{min} : batas nilai minimum baru

χ -Sim co-similarity

χ -Sim adalah ukuran berbasis *co-similarity* yang membentuk matriks similaritas siswa (SS) dan similaritas pelajaran (SP) secara iteratif, ditambahkan dengan L_k normalisasi (jarak minkowski), normalisasi pseudo dengan parameter $\lambda < 1$ yang bertujuan $\forall a, b \in 1..r, ss_{ab} \in [0,1]$ dimana ss_{ab} merupakan elemen matriks similaritas siswa, misalkan $I^{\lambda} = \left((i_{ab})^{\lambda} \right)_{a,b}$ dimana I^{λ} merupakan data matriks *Try Out* siswa dan $\frac{\otimes}{\otimes} I^{\lambda} \frac{\otimes^T}{\otimes}$ merupakan *transpose* data matriks *Try Out* siswa, masing-masing ditambahkan dengan L_k normalisasi. (Hussain, 2010). Berikut langkah-langkah dari χ -Sim co-similarity (Hussain, Bisson, & Grimal, 2010):

1. Menginisialisasi matriks similaritas baris $SS^{(0)}$ (siswa) dan $SP^{(0)}$ (pelajaran) dengan matriks identitas.
2. Tiap iterasi t , kita menghitung matriks similaritas antar siswa $SS^{(t)}$ yang baru dengan menggunakan matriks similaritas pelajaran $SP^{(t-1)}$:

$$SS^{(t)} = I^{\lambda} \times SP^{(t-1)} \times \left(I^{\lambda} \right)^T \quad (2)$$

$$\text{dan } \forall a, b \in 1..k, SS_{ab}^{(t)} = \frac{\sqrt[\lambda]{SS_{ab}^{(t)}}}{2\sqrt[\lambda]{SS_{aa}^{(t)} \times SS_{bb}^{(t)}}} \quad (3)$$

kita lakukan hal yang sama untuk matriks similaritas pelajaran $SP^{(t)}$:

$$SP^{(t)} = \left(I o^\lambda \right)^T x SS^{(t-1)} x I o^\lambda \quad (4)$$

dan $\forall a, b \in 1, \dots, k, SS_{ab}^{(t)} = \frac{\sqrt[\lambda]{SP_{ab}^{(t)}}}{2\lambda \sqrt[\lambda]{SP_{aa}^{(t)} x SP_{bb}^{(t)}}} \quad (5)$

- Langkah kedua kita ulangi sampai 3 atau 4 kali, $t = 3$ atau 4 kali cukup untuk mengupdate matriks $SS^{(0)}$ (siswa) dan $SP^{(0)}$ (pelajaran) secara iteratif sesuai yang disarankan oleh peneliti sebelumnya.

Algoritma K-means Clustering

Salah satu algoritma pengelompokan yang populer yaitu *k-means clustering*. Diberikan himpunan observasi (x_1, x_2, \dots, x_n) dimana tiap observasi merupakan sebuah *d-dimensional real vector*, *k-means* bertujuan untuk mempartisi n observasi menjadi k himpunan ($k \leq n$) $S = \{S_1, S_2, \dots, S_k\}$ untuk meminimalkan jumlah kuadrat dalam *cluster*

$$\arg \min_s \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \quad (6)$$

dimana μ_i adalah rata-rata poin dalam S_i (Mohamad & Usman, 2013). Berikut merupakan langkah dari algoritma *k-means* (Kaufman & Rousseeuw, 2005):

- Tentukan jumlah *cluster* (k), tetapkan pusat *cluster* (*centroid*) secara acak.
- Hitung jarak setiap objek ke pusat *clusters* menggunakan *Euclidean distance* dengan rumus berikut

$$d_{ik} = \sqrt{\sum_{i,k=1}^n (x_i - c_k)^2} \quad (7)$$

Keterangan:

- d_{ik} = jarak objek i ke pusat kelompok k
- n = dimensi data
- x_i = koordinat dari objek i
- c_k = koordinat pusat data (*centroid*)

- Kelompokkan data ke dalam *cluster* yang memiliki jarak dengan pusat *cluster* yang paling dekat.
- Hitung pusat *cluster* baru dengan rumus berikut.

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \quad (8)$$

Keterangan:

- μ_k = titik *centroid* dari *cluster* ke- k
- N_k = banyaknya data pada *cluster* ke- k

x_q = data ke- q pada *cluster* ke- k

5. Ulangi langkah 2 sampai dengan 4 hingga sudah tidak ada lagi objek yang berpindah ke *cluster* lain.

Metode Indeks *Silhouette*

Indeks *Silhouette* merupakan suatu ukuran statistik yang digunakan untuk menyeleksi permasalahan penentuan jumlah *cluster* optimal atau seberapa baik setiap objek terletak dalam *cluster* (Nicolaus, Sulistianingsih, & Perdana, 2016). *Silhouette* dapat digunakan untuk semua jenis data, dan struktur pengelompokan apapun (Zaki & Meira, Jr, 2018). Indeks *Silhouette* terletak antara nilai -1 dan 1 dimana semakin mendekati 1 berarti data tersebut berada pada kelompok yang tepat, sebaliknya jika semakin mendekati -1 berarti data tersebut berada di kelompok yang salah (Starczewski, 2017). Berikut ini merupakan langkah-langkah untuk menghitung indeks *silhouette* (Kaufman & Rousseeuw, 2005):

1. Berdasarkan persamaan (9), untuk setiap objek i hitung rata-rata jarak dari objek i dengan seluruh objek yang berada dalam *cluster* yang sama sebut $a(i)$ dan hitung rata-rata jarak dari objek i dengan objek-objek yang berada pada *cluster* yang lainnya sebut $b(i)$, kemudian hitung nilai *silhouette* objek yaitu $sil(x_i)$.

$$sil(x_i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (9)$$

2. Untuk setiap *cluster* ke- j , hitung rata-rata dari *cluster* objek *silhouette* dengan menggunakan persamaan (10).

$$sil(C_j) = \frac{1}{|C_j|} \sum_{x_i \in C_j} sil(x_i) \quad (10)$$

3. Hitung indeks *silhouette* dengan menggunakan persamaan (11) dimana k adalah banyaknya *cluster*.

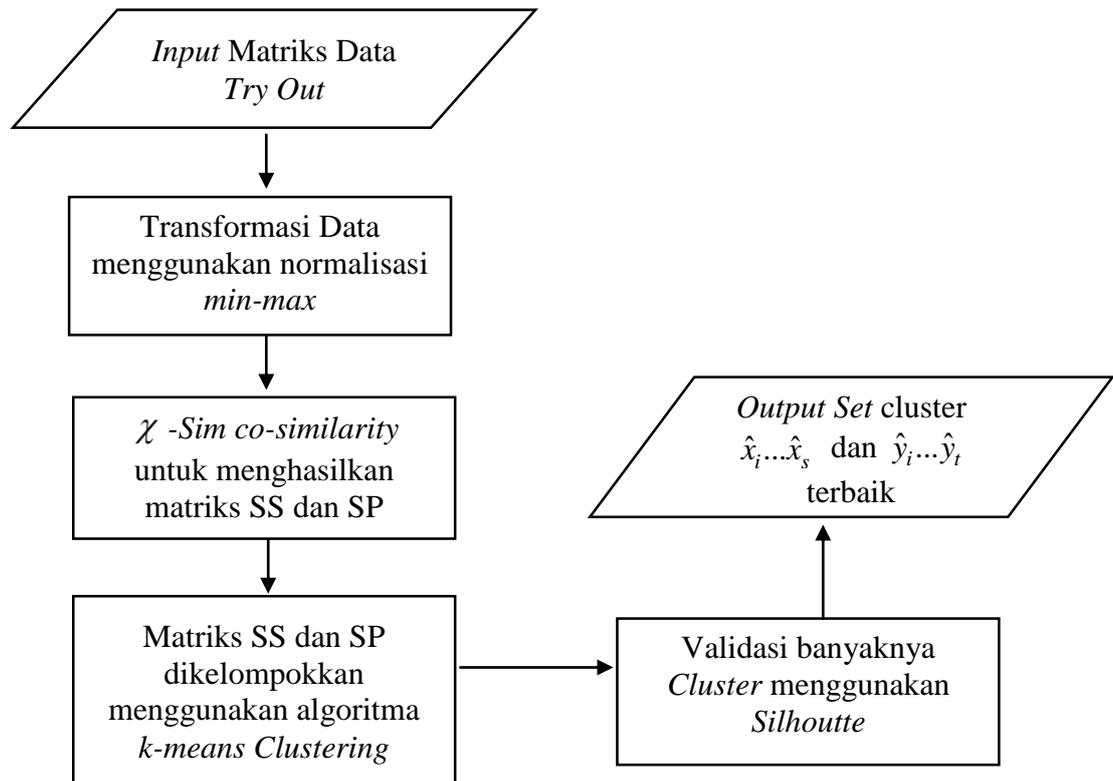
$$sil(k) = \frac{1}{|k|} \sum_{r=1}^k sil(c_r) \quad (11)$$

Metode

Metode penelitian yang digunakan dalam penelitian ini adalah studi literatur, pengumpulan data, pembuatan program komputer dari algoritma yang diajukan di perangkat lunak R versi 3.6.1, analisis terhadap hasil pengelompokan, dan penarikan kesimpulan. Data yang digunakan pada penelitian ini merupakan data sekunder yang diambil dari data hasil *Try Out* siswa kelas IX MTs Alhuda Gorontalo Tahun Ajaran 2018/2019 dengan banyaknya siswa 164 orang.

Untuk mendapatkan *output*, diperlukan tahapan-tahapan. Langkah pertama yaitu mentransformasi data dengan menggunakan metode normalisasi min-max untuk mentransformasi matriks X menjadi matriks X_n . Selanjutnya, diterapkan metode χ -*Sim co-similarity* untuk menghasilkan matriks similaritas siswa (SS) dan similaritas pelajaran (SP). Masing-masing SS dan SP kemudian dikelompokkan dalam s dan t kelompok menggunakan algoritma *k-means clustering* sebagai $\hat{x}_1 \dots \hat{x}_s$ dan $\hat{y}_1 \dots \hat{y}_t$ dan menggunakan metode *Silhouette* untuk menentukan kelompok yang terbaik, yang kemudian ke penarikan kesimpulan. *Flowchart* atau diagram alir dapat

dilihat pada Gambar 1.



Gambar 1. Flowchart analisis data

HASIL DAN PEMBAHASAN

Pada Gambar 2, diberikan matriks data hasil Try Out Siswa. Pada matriks tersebut, baris-baris merepresentasikan siswa dan kolom-kolom merepresentasikan mata pelajaran.

	Bahasa Indonesia	Bahasa Inggris	Matematika	IPA
1	82	62	72.5	72.5
2	70	56	72.5	57.5
3	64	56	42.5	55.0
4	78	60	72.5	55.0
5	64	64	70.0	57.5
6	62	66	75.0	55.0
7	64	70	80.0	62.5
8	64	54	70.0	55.0
9	70	56	72.5	52.5
10	68	64	67.5	55.0
11	88	80	67.5	60.0

Gambar 2. Matriks data hasil Try Out Siswa

Dalam pencarian *Set cluster* $\hat{x}_1 \dots \hat{x}_s$ dan $\hat{y}_1 \dots \hat{y}_t$ terbaik, dimulai dengan mentransformasi matriks data hasil *Try Out* Siswa. Siswa pada baris pertama dinamakan siswa 1 dan siswa pada baris kedua dinamakan siswa 2. Tabel 1 menampilkan data siswa 1 dan siswa 2 setelah dinormalisasi.

Tabel 1. Hasil Transformasi Data (Normalisasi) pada siswa 1 dan siswa 2

Siswa 1				Siswa 2			
Bahasa Indonesia	Bahasa Inggris	Mat	IPA	Bahasa Indonesia	Bahasa Inggris	Mat	IPA
1	0	0,525	0,525	0,848485	0	1	0,090909

Setelah mentransformasi data (normalisasi), langkah selanjutnya adalah membentuk matriks similaritas siswa dan similaritas pelajaran. Matriks similaritas yang digunakan dalam penelitian ini adalah matriks yang diperoleh dari metode χ -Sim co-similarity dengan nilai $\lambda = \frac{1}{50}$ dan terhenti pada iterasi 4. Berikut merupakan matriks similaritas.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	1.0000000	0.9976819	0.9929626	0.9984687	0.9980623	0.9977974	0.9961894	0.9966388	0.9975869	0.9987655	0.9988368
2	0.9976819	1.0000000	0.9831040	0.9994528	0.9993213	0.9993384	0.9993887	0.9999024	0.9997454	0.9987693	0.9958732
3	0.9929626	0.9831040	1.0000000	0.9871114	0.9865269	0.9858491	0.9814700	0.9804859	0.9841584	0.9893812	0.9947507
4	0.9984687	0.9994528	0.9871114	1.0000000	0.9999675	0.9999327	0.9994455	0.9990496	0.9998421	0.9998509	0.9981041
5	0.9980623	0.9993213	0.9865269	0.9999675	1.0000000	0.9999906	0.9995900	0.9989668	0.9998520	0.9998257	0.9979756
6	0.9977974	0.9993384	0.9858491	0.9999327	0.9999906	1.0000000	0.9996989	0.9990404	0.9998844	0.9997424	0.9977236
7	0.9961894	0.9993887	0.9814700	0.9994455	0.9995900	0.9996989	1.0000000	0.9994332	0.9998281	0.9988848	0.9957792
8	0.9966388	0.9999024	0.9804859	0.9990496	0.9989668	0.9990404	0.9994332	1.0000000	0.9995876	0.9981549	0.9946408
9	0.9975869	0.9997454	0.9841584	0.9998421	0.9998520	0.9998844	0.9998281	0.9995876	1.0000000	0.9994051	0.9968597
10	0.9987655	0.9987693	0.9893812	0.9998509	0.9998257	0.9997424	0.9988848	0.9981549	0.9994051	1.0000000	0.9989855
11	0.9988368	0.9958732	0.9947507	0.9981041	0.9979756	0.9977236	0.9957792	0.9946408	0.9968597	0.9989855	1.0000000

Gambar 3. Matriks Similaritas Siswa

	V1	V2	V3	V4
1	1.0000000	0.9989300	0.9902756	0.9981104
2	0.9989300	1.0000000	0.9837760	0.9988501
3	0.9902756	0.9837760	1.0000000	0.9803574
4	0.9981104	0.9988501	0.9803574	1.0000000

Gambar 4. Matriks Similaritas Pelajaran

Pada Gambar 3 dan Gambar 4, entri-entri dari matriks similaritas konvergen ke suatu nilai dan membatasi nilai similaritas dengan interval $[0,1]$. Sesuai dengan sifat matriks similaritas, $I = I^T$.

Matriks similaritas siswa dan similaritas pelajaran yang diperoleh pada metode χ -Sim co-similarity dikelompokkan menggunakan algoritma *k-means clustering*. Untuk menentukan banyaknya kelompok terbaik digunakan *Silhouette*. Untuk menganalisis lebih lanjut hasil pengelompokan, akan dibandingkan hasil pengelompokan matriks similaritas siswa dan similaritas pelajaran pada iterasi tiga dan empat.

Tabel 2 berikut menampilkan nilai *Silhouette* untuk beberapa kemungkinan banyaknya *cluster* pada matriks similaritas siswa data nilai *Try Out* untuk $k = 2, 3, \dots, 10$.

Tabel 2. Nilai *Silhouette* pada Matriks Similaritas Siswa Data Nilai *Try Out*

Iterasi 3		Iterasi 4	
Banyak kelompok	Nilai <i>Silhouette</i>	Banyaknya kelompok	Nilai <i>Silhouette</i>
2	0.6182013	2	0.7491923
3	0.5818186	3	0.9136687
4	0.5972806	4	0.9755781
5	0.5654646	5	0.7887290
6	0.5890485	6	0.6418580
7	0.3771929	7	0.6355246
8	0.4035226	8	0.6484256
9	0.3936987	9	0.5615189
10	0.3886333	10	0.5921562

Berdasarkan Tabel 2, diperoleh *Silhouette* terbesar yaitu 0,9755781 pada iterasi keempat yang mempartisi data menjadi 4 *cluster*. Hal ini berarti jarak *Euclidean* antar siswa pada masing-masing kelompok dengan gabungan metode χ -Sim co-similarity (iterasi 4) dan *k-means clustering* lebih dekat dibandingkan dengan *k-means clustering* dan χ -Sim co-similarity (iterasi 3). Oleh karena itu, gabungan metode χ -Sim co-similarity (iterasi 4) dan *k-means clustering* memberikan hasil pengelompokan yang lebih homogen pada setiap kelompoknya dibandingkan metode *k-means* dan *chi-sim co-similarity* (iterasi 3) untuk data *Try Out* Siswa kelas IX.

Tabel 3 berikut menampilkan nilai *Silhouette* untuk $k = 2$ dan 3 pada matriks similaritas pelajaran data *Try Out* siswa kelas IX. Dalam kasus ini, akan dibandingkan hasil pengelompokan matriks similaritas pelajaran yang dihasilkan pada iterasi ketiga dan matriks similaritas pelajaran yang dihasilkan pada iterasi keempat.

Tabel 3. Nilai *Silhouette* pada Matriks Similaritas Pelajaran Data Nilai *Try Out*

Iterasi 3		Iterasi 4	
Banyak kelompok	Nilai <i>Silhouette</i>	Banyak kelompok	Nilai <i>Silhouette</i>
2	0,5054545	2	0,5756133
3	0,06332955	3	0,2594635

Berdasarkan Tabel 3, diperoleh nilai *silhouette* terbesar yaitu **0,5756133** pada iterasi keempat yang mempartisi data menjadi 2 *cluster*. Hal ini berarti jarak *Euclidean* antar pelajaran pada masing-masing kelompok dengan gabungan metode χ -*Sim co-similarity* (iterasi 4) dan *k-means clustering* lebih dekat dibandingkan dengan *k-means clustering* dan χ -*sim co-similarity* (iterasi 3). Oleh karena itu, gabungan metode χ -*Sim co-similarity* (iterasi 4) dan *k-means clustering* memberikan hasil pengelompokan yang lebih homogen pada setiap kelompoknya dibandingkan metode *k-means* dan χ -*Sim co-similarity* (iterasi 3) untuk data *Try Out* Siswa kelas IX.

Tabel 4 menampilkan informasi hasil dari *cluster* yang terbentuk pada matriks similaritas siswa. Ada sebanyak 67 siswa pada *cluster* 1, 9 siswa pada *cluster* 2, 45 siswa pada *cluster* 3 dan 43 siswa pada *cluster* 4.

Tabel 4. Informasi hasil *cluster* matriks similaritas siswa

Anggota Cluster 1	Anggota Cluster 2	Anggota Cluster 3	Anggota Cluster 4
3, 12, 15, 16, 18, 21, 22, 23, 27, 28, 29, 30, 31, 34, 36, 37, 38, 39, 41, 43, 44, 46, 47, 52, 57, 58, 59, 62, 64, 65, 68, 69, 70, 76, 79, 80, 82, 86, 88, 90, 91, 92, 97, 98, 102, 103, 110, 114, 117, 122, 123, 124, 125, 126, 128, 131, 134, 135, 136, 139, 148, 151, 156, 157, 159, 160, dan 164.	13, 14, 40, 93, 99, 113, 129, 133, dan 143.	1, 2, 4, 5, 6, 7, 8,9, 10, 11, 19, 25, 26, 32, 35, 53, 55, 56, 71, 72, 73, 74, 75, 77, 78, 81, 84, 85, 87, 89, 95, 96, 100, 107, 115, 116, 118, 120, 130, 132, 137, 138, 140, 141, 142, 144, 146, 149, 152, 153, 161, dan 163.	17, 20, 24, 33, 42, 45, 48, 49, 50, 51, 54, 60, 61, 63, 66, 67, 83, 94, 101, 104, 105, 106, 108, 109, 111, 112, 119, 121, 127, 145, 147, 150, 154, 155, 158, dan 162.

Berdasarkan Tabel 4, anggota-anggota pada *cluster* 1 terdiri dari siswa-siswa yang memiliki kinerja buruk pada mata pelajaran Bahasa Inggris, Matematika dan IPA. *Cluster* 2 terdiri dari siswa-siswa yang memiliki kinerja buruk pada mata pelajaran Bahasa Inggris dan Matematika. *Cluster* 3 terdiri dari siswa yang memiliki kinerja baik pada semua pelajaran dan *cluster* 4 terdiri dari siswa-siswa yang memiliki kinerja buruk pada mata pelajaran IPA.

Tabel 5. Informasi hasil *cluster* matriks similaritas pelajaran

Anggota Cluster 1	Anggota Cluster 2
Bahasa Indonesia dan Bahasa Inggris	Matematika dan IPA

Tabel 5 menampilkan informasi hasil dari *cluster* yang terbentuk pada matriks similaritas pelajaran. Ada sebanyak 2 mata pelajaran pada *cluster* 1 dan 2 mata pelajaran pada *cluster* 2. Hal ini berarti karakteristik *cluster* 1 dan *cluster* 2 berbeda. Mata pelajaran Bahasa Indonesia dan Bahasa Inggris memiliki rata-rata diatas

sedangkan mata pelajaran matematika dan IPA memiliki rata-rata rendah.

KESIMPULAN DAN SARAN

Pada penelitian ini, χ -sim co-similarity dan algoritma *k-means clustering* dapat diterapkan pada data hasil *Try Out* Siswa dalam mengelompokkan siswa-siswa dan mata pelajaran. Pada pengelompokkan SS diperoleh nilai *Silhouette* terbesar yaitu 0,9755781 pada iterasi keempat yang mempartisi menjadi 4 *cluster* sebagai berikut sebanyak 67 siswa pada *cluster* 1, 9 siswa pada *cluster* 2, 45 siswa pada *cluster* 3 dan 43 siswa pada *cluster* 4. anggota-anggota pada *cluster* 1 terdiri dari siswa-siswa yang memiliki kinerja buruk pada mata pelajaran Bahasa Inggris, matematika dan IPA. *Cluster* 2 terdiri dari siswa-siswa yang memiliki kinerja buruk pada mata pelajaran Bahasa Inggris dan Matematika. *Cluster* 3 terdiri dari siswa yang memiliki kinerja baik pada semua pelajaran dan *cluster* 4 terdiri dari siswa-siswa yang memiliki kinerja buruk pada mata pelajaran IPA.

Pada pengelompokkan SP diperoleh nilai *Silhouette* terbesar yaitu 0,5756133 pada iterasi keempat yang mempartisi menjadi 2 *cluster* sebagai berikut Pada *cluster* 1 terdiri dari Bahasa Indonesia dan Bahasa Inggris dan pada kelompok 2 terdiri dari Matematika dan IPA. Mata pelajaran Bahasa Indonesia dan Bahasa Inggris memiliki rata-rata diatas sedangkan mata pelajaran Matematika dan IPA memiliki rata-rata dibawah.

Beberapa saran yang dapat diambil dari penelitian ini dan dikembangkan untuk penelitian selanjutnya yaitu dapat menggunakan algoritma *clustering* yang lainnya atau algoritma *biclustering* yang ada sehingga bisa mengidentifikasi tingkat kinerja siswa dalam mata pelajaran tertentu saja.

DAFTAR PUSTAKA

- Bergeron, B. (2002). *Bioinformatics Computing*. In *Prentice Hall PTR* (Vol. 36). <https://doi.org/10.1109/IJCBS.2009.8>
- Bustamam, A., Zubedi, F., & Siswantining, T. (2018). Implementation χ -sim co-similarity and agglomerative hierarchical to cluster gene expression data of lymphoma by gene and condition. *AIP Conference Proceedings*, 2023. <https://doi.org/10.1063/1.5064218>
- Chin, A. J., Mirzal, A., & Harona, H. (2015). *Review FS clustering*. 8, 1–5.
- Han, J. (2011). *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Retrieved from <http://www.amazon.co.uk/Data-Mining-Concepts-Techniques-Management/dp/0123814790>
- Hussain, S. F., Bisson, G., & Grimal, C. (2010). An improved co-similarity measure for document clustering. *Proceedings - 9th International Conference on Machine Learning and Applications, ICMLA 2010*, 190–197. <https://doi.org/10.1109/ICMLA.2010.35>
- Hussain, S. F., & Ramazan, M. (2016). Biclustering of human cancer microarray data using co-similarity based co-clustering. *Expert Systems with Applications*, 55(February 2016), 520–531. <https://doi.org/10.1016/j.eswa.2016.02.029>
- Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data : an introduction to cluster analysis / Leonard Kaufman, Peter J. Rousseeuw*. Retrieved from <http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cab00362a&AN=neos.5309634&site=eds->

-
- live&scope=site;%5Cnhttp://www.loc.gov/catdir/enhancements/fy0626/2005278659-t.html;%5Cnhttp://www.loc.gov/catdir/e
- Liu, B., Xin, Y., Cheung, R. C. C., & Yan, H. (2014). GPU-based biclustering for microarray data analysis in neurocomputing. *Neurocomputing*, 134, 239–246. <https://doi.org/10.1016/j.neucom.2013.06.049>
- Mohamad, I. Bin, & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6(17), 3299–3303. <https://doi.org/10.19026/rjaset.6.3638>
- Mondal, B., & Paul Choudhury, J. (2013). A Comparative Study on K Means and PAM Algorithm using Physical Characters of Different Varieties of Mango in India. *International Journal of Computer Applications*, 78(5), 21–24. <https://doi.org/10.5120/13485-1189>
- Nicolaus, Sulistianingsih, E., & Perdana, H. (2016). Penentuan jumlah cluster optimal pada median linkage dengan indeks validitas silhouette menggunakan himpunan variabel yang ditentukan untuk mengelompokkan objek. *Buletin Ilmiah Math. Stat. Dan Terapannya (Bimaster)*, 05(2), 97–102.
- Patro, S. G. K., & sahu, K. K. (2015). Normalization: A Preprocessing Stage. *Iarjset*, 20–22. <https://doi.org/10.17148/iarjset.2015.2305>
- Starzewski, A. (2017). A new validity index for crisp clusters. *Pattern Analysis and Applications*, 20(3), 687–700. <https://doi.org/10.1007/s10044-015-0525-8>
- X. Wu and VKumar. (2009). *Top 10 Algorithms in Data Mining*. Chapman & Hall/CRC.
- Zaki, M. J., & Meira, Jr, W. (2018). Clustering Validation. In *Data Mining and Analysis*. <https://doi.org/10.1017/cbo9780511810114.018>