Jurnal Teknologi Terpadu

https://journal.nurulfikri.ac.id/index.php/jtt
ISSN : 2477-0043 ISSN ONLINE : 2460-7908

PENERAPAN *K-MEANS* DAN *RANK ORDER CENTROID* PADA PROPORSI INDIVIDU DENGAN KETERAMPILAN TEKNOLOGI INFORMASI DAN KOMPUTER

Diana Nurfitriana[1], Apriade Voutama[2]

[1,2] Sistem Informasi, Universitas Singaperbangsa Karawang
Karawang, Jawa Barat, Indonesia 41361
diana.nurfitriana19011@student.unsika.ac.id, apriade.voutama@staff.unsika.ac.id

Abstract

Technological developments occur so quickly, resulting in continuous changes that qualified human resources are needed to support the endless times that run. This study will classify individuals with information technology and computer skills in Indonesia based on region. This research used K-Means clustering, the Rank Order Centroid method, and the Davies-Bouldin Index clustering evaluation method to assess accuracy. K-means clustering is a simple algorithm and does not require a target class. There are areas for improvement in the K-Means process, namely at the initial centroid determination stage. Therefore, the ROC method is used. Based on data taken from the website of Badan Pusat Statistik Nasional about the proportion of productive age individuals 15-59 years who have Information and Computer Technology skills by the province during 2017-2021. It produces 3 clusters, including a high-level cluster in which there are 8 provinces, a medium-level cluster in which there are 22 provinces, and a low-level cluster in which there are 4 provinces, and obtained a DBI value of 0.163625 which is close to 0, meaning that the quality of the accuracy of the clustering results is good. Based on clustering results with good accuracy, using K-Means can be combined with ROC and is quite effective. The government can use the results of this study to prioritize improving the quality of human resources in areas with low-level information and computer technology skills. Suggestions for further research using other clustering algorithms and ROC as a comparison.

Keywords: *data mining, Davies-Bouldin Index, information and computer technology, K-Means clustering, Rank Order Centroid*

Abstrak

Perkembangan teknologi yang berlangsung begitu cepat mengakibatkan perubahan yang terus terjadi dan sumber daya manusia yang mumpuni dibutuhkan guna mendukung zaman yang kian berkembang. Penelitian ini bertujuan untuk mengelompokkan individu keterampilan teknologi informasi dan komputer di Indonesia berdasarkan wilayah. Penelitian ini dilakukan dengan menggunakan *K-Means clustering* dan metode *Rank Order Centroid*, serta metode evaluasi *clustering Davies-Bouldin Index* untuk menilai akurasi. *K-means clustering* merupakan algoritma yang sederhana dan tidak membutuhkan target kelas. Terdapat kekurangan pada proses *K-Means* yaitu pada tahap penentuan *centroid* awal, maka dari itu digunakan metode ROC. Berdasarkan data yang diambil dari situs Badan Pusat Statistik Nasional tentang data proporsi individu usia 15-59 tahun dengan keterampilan TIK menurut provinsi selama rentang tahun 2017-2021 menghasilkan 3 *cluster* di antaranya adalah *cluster* tingkat tinggi terdapat 8 provinsi, *cluster* tingkat sedang terdapat 22 provinsi dan *cluster* tingkat rendah terdapat 4 provinsi dan didapatkan nilai evaluasi *DBI* sebesar 0,163625 yang mendekati 0, berarti kualitas akurasi dari hasil *clustering* baik. Berdasarkan hasil *clustering* dengan akurasi yang baik, penggunaan *K-Means* dapat dikombinasikan dengan *ROC* dan cukup efektif. Dari hasil penelitian ini dapat dimanfaatkan oleh pemerintah untuk meningkatkan kualitas sumber daya manusia di wilayah dengan tingkat keterampilan teknologi informasi dan komputer yang rendah. Saran untuk penelitian selanjutnya, menggunakan algoritma *clustering* lain dan *ROC* sebagai perbandingan.

Kata kunci: *data mining, Davies-Bouldin Index, K-Means clustering, Rank Order Centroid*, teknologi informasi dan komputer

1. PENDAHULUAN

Seiring berkembangnya teknologi pada era revolusi industri 4.0, teknologi informasi dan komputer menjadi tolak ukur penting dalam perkembangan teknologi yang semakin canggih dan efisien. Teknologi yang saat ini berkembang pesat mendorong pemanfaatan teknologi ini di segala

Diterima 7 November 2022, Direvisi 28 Mei 2023, Diterima untuk publikasi 13 Juni 2023

bidang dan dapat membawa keuntungan untuk pemrosesan dan pengambilan informasi untuk meningkatkan kualitas bisnis di masa depan [1]. Pemanfaatan teknologi ini dapat mempermudah proses kegiatan baik dari segi lokasi maupun biaya [2]. Perkembangan teknologi yang berlangsung begitu cepat mengakibatkan perubahan yang terus terjadi dan sumber daya manusia yang mumpuni dibutuhkan untuk mendukung perkembangan zaman yang terus berjalan tanpa henti. Tidak menutup kemungkinan bahwa manusia dituntut untuk mengembangkan keterampilan terutama di bidang ini untuk menyesuaikan dengan pekerjaan baru yang tidak pernah terpikirkan sebelumnya. Pekerjaan yang dimaksud beberapa di antaranya adalah *social media specialist, content creator, cyber security, data scientist, data engineer* dan *big data specialist*. Dilihat dari hal tersebut, keterampilan teknologi informasi dan komputer kini telah menjadi kebutuhan primer dan bukan sekunder [3]. Untuk mencapai sumber daya manusia yang berkualitas diperlukan pendidikan yang berkualitas pula. Dari segi proses pembelajaran dan juga tenaga pendidik yang bermutu. Oleh karena itu, penelitian ini perlu dilakukan untuk memberikan informasi kepada pemerintah mengenai daerah-daerah di Indonesia yang memiliki tingkat keterampilan teknologi informasi dan komputer yang rendah dalam menghadapi kemajuan teknologi, supaya tidak menjadi masyarakat yang tertinggal.

Data yang berkaitan dengan permasalahan ini bersumber dari situs http://bps.go.id. Data tersebut merupakan salah satu Indikator Pembangunan Berkelanjutan bagian Pendidikan Berkualitas, yaitu proporsi remaja dan dewasa usia (15-59 tahun) dengan keterampilan Teknologi Informasi dan Komputer (TIK) menurut provinsi. Rentang waktu data yang diteliti adalah tahun 2017-2021. Untuk mengetahui daerah-daerah yang memiliki keterampilan teknologi informasi dan komputer yang rendah perlu dilakukan pengelompokan pada keseluruhan wilayah di Indonesia menjadi kelompok tingkat rendah, tingkat sedang dan tingkat tinggi.

Terdapat banyak metode di bidang teknologi informasi dalam mengolah data tersebut, salah satunya adalah *data mining*. *Data mining* adalah suatu proses ekstraksi pola dari data dan informasi berukuran besar yang menghasilkan pengetahuan untuk menyederhanakan data dan mendapatkan informasi yang informatif serta bermanfaat dengan bantuan ilmu statistik dan matematika [4]. *Data mining* dapat digunakan di berbagai bidang dan untuk tujuan yang berbeda yaitu untuk meningkatkan pengetahuan, meningkatkan penjualan di beberapa sektor, dll [5]. Algoritma *K-Means* adalah metode umum dan paling sederhana dalam *clustering. Clustering* adalah termasuk dari ilmu *data mining* tanpa arahan (*unsupervised learning*). *Clustering* atau klasterisasi adalah proses membagi data ke dalam kelas atau klaster berdasarkan kesamaan. *Clustering* menempatkan data yang kadar kesamaannya tinggi pada klaster yang sama sedangkan data yang kadar kesamaannya rendah dimasukkan dalam klaster yang berbeda. Dalam algoritma *K-Means Clustering*, *K* berarti konstanta untuk jumlah klaster yang dibutuhkan dan *Means* berarti nilai rata-rata dari sekelompok data yang dalam hal ini didefinisikan sebagai klaster. Kemiripan anggota klaster diukur dengan kedekatan objek dengan *Mean* dalam klaster atau disebut sebagai *centroid.* Jadi, *K-Means Clustering* adalah teknik analisis data dengan proses pemodelan tanpa arahan (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi [6]. Alasan pemilihan algoritma *K-Means* ini karena dalam mengelompokkan data tidak perlu mengetahui target kelas dan merupakan metode yang sederhana dibanding *clustering* lainnya [7].

Proses dari algoritma *K-Means* dimulai dari tahap pemilihan jumlah klaster yang diinginkan, dan selanjutnya menentukan *centroid* awal secara acak sejumlah dengan klaster yang telah ditentukan sebelumnya. Setelah itu, jarak *Euclidean* dari setiap data ke pusat klaster dihitung. Untuk menghindari pergeseran data ke klaster lain, proses ini dilakukan secara berulang [8]. Terdapat celah kekurangan pada proses *K-Means* yaitu pada tahap penentuan *centroid* awal yang secara acak, karena berpengaruh terhadap hasil *clustering* yang berbeda dan tidak konsisten. Maka dari itu, penentuan *centroid* awal pada penelitian ini akan menggunakan metode *Rank Order Centroid* (*ROC*). *ROC* merupakan metode yang mementingkan tingkat prioritas dari kriteria, teknik *ROC* melakukan pembobotan pada setiap kriteria berdasarkan *ranking* prioritas. Penentuan prioritas dilakukan dengan cara mencari nilai tertinggi dan menjadikannya nilai yang paling penting dibandingkan nilai lainnya [7]. Metode yang akan digunakan untuk evaluasi hasil *clustering* adalah *Davies-Bouldin Index (DBI).* Evaluasi dengan *DBI* mempunyai skema evaluasi klaster internal yang menentukan baik atau tidaknya hasil klaster dilihat dari kuantitas dan kedekatan antara hasil klaster [9]. Pendekatan dalam pengujian nilai *DBI* mencakup nilai separasi dan kohesi. Klaster dapat dikatakan optimal apabila klaster tersebut mempunyai nilai kohesi yang rendah dan nilai separasi yang tinggi [10].

Beberapa penelitian yang mendasari penelitian penulis adalah sebagai berikut. Penelitian sebelumnya yang dilakukan oleh [11] menunjukkan bahwa *K-Means clustering* dapat digunakan untuk mengelompokkan data wilayah (puskesmas) di Banyuwangi yang dikelompokkan menjadi 3 kategori. Hasil dari penelitian tersebut menjadi informasi bagi Dinas Kesehatan dalam meningkatkan kinerja di puskesmas dengan target imunisasi kurang. Namun, pada penelitian ini akurasi dari hasil *clustering* tidak diketahui. Pada penelitian lainnya oleh [12] dalam proses algoritma *K-Means* pada penentuan jumlah *cluster* menggunakan metode *elbow* dengan hasil 3 *cluster* optimal dan evaluasi menggunakan *Silhouette Coefficient* dengan nilai 0,4960 yang menunjukkan bahwa kualitas *cluster* yang baik. Penelitian ini menggunakan *dataset* Penyakit ISPA di kabupaten Karawang dari tahun 2019-2021. Pada penelitian oleh [7] membahas tentang perbandingan kinerja algoritma

K-Means konvensional dengan *K-Means* menggunakan metode *ROC* untuk penentuan *centroid*. Dikatakan bahwa *centroid* awal pada algoritma *K-Means* sangat berpengaruh pada kualitas *cluster* yang dihasilkan. Hasil penelitian menunjukkan bahwa nilai akurasi dari *K-Means* dengan metode *ROC* memiliki akurasi yang lebih tinggi dibanding *K-Means* konvensional. Maka dari itu, penelitian sekarang menggunakan metode *ROC* dalam *K-Means* untuk mendapatkan nilai akurasi yang lebih baik. Perbedaan penelitian sekarang dengan penelitian [7] yaitu pada evaluasi hasil *clustering*. Penelitian tersebut melakukan evaluasi *clustering* menggunakan *Relative Standard Deviation* (*RSD*) [7]. Penelitian sekarang akan menggunakan evaluasi *Davies-Bouldin Index*, berbeda dengan penelitian [7] dan [12].

2. METODE PENELITIAN

2.1 Metode pengumpulan data dan metode pengujian

Penelitian ini menggunakan data tentang proporsi individu usia produktif (15-59 tahun) dengan keterampilan Teknologi Informasi dan Komputer yang diperoleh dari situs resmi terbuka milik lembaga pemerintah Badan Pusat Statistik Nasional https://www.bps.go.id. Data yang digunakan adalah data dalam rentang waktu tahun 2017-2021 yang terdiri dari 34 provinsi. *Tools* yang digunakan untuk membantu perhitungan data di penelitian ini adalah Microsoft Excel.

Algoritma yang diterapkan pada penelitian ini adalah metode *K-Means clustering* untuk menentukan tiap-tiap *cluster*nya dan metode *Rank Order Centroid (ROC)* untuk menentukan *centroid* awal. Metode yang digunakan dalam evaluasi *cluster* yang telah ditentukan menggunakan *Davies-Bouldin Index (DBI).*

2.2 Tahapan penelitian

Tahapan penelitian merupakan rentetan proses yang sistematis selama dilakukannya penelitian. Tahapan penelitian dimaksudkan untuk memudahkan pencapaian hasil penelitian, menyelesaikan penelitian tepat waktu, dan memajukan penelitian sesuai dengan yang diharapkan [13]. Tahapan penelitian ini tertera pada gambar 1.

Gambar 1. Tahapan Penelitian

2.2.1 Identifikasi Masalah

Identifikasi masalah merupakan bagian dari proses paling awal penelitian atau dapat dipahami sebagai langkah pertama dalam penelitian sebagai mendefinisikan masalah dan mencoba untuk membuat definisi tersebut menjadi lebih terukur. Masalah pada penelitian ini yaitu tidak adanya pengelompokan tingkatan wilayah yang memiliki keterampilan Teknologi Informasi dan Komputer (TIK) rendah di Indonesia.

2.2.2 Studi Literatur

Studi literatur dilakukan dengan mencari referensi jurnal penelitian/tulisan penelitian yang berkaitan dengan metode-metode yang dapat digunakan pada penelitian ini. Dalam hal ini yaitu, algoritma *K-Means clustering*, metode *ROC* dan pengujian *DBI*.

2.2.3 Pengumpulan Data

Pada penelitian ini data dikumpulkan dari situs resmi terbuka milik lembaga pemerintah Badan Pusat Statistik Nasional https://www.bps.go.id. Data tersebut merupakan salah satu data Indikator Pembangunan Berkelanjutan bagian Pendidikan Berkualitas yaitu proporsi remaja dan dewasa usia (15-59 tahun) dengan keterampilan Teknologi Informasi dan Komputer menurut provinsi. Data berikut hasil dari survei pada proporsi remaja (umur 15-24 tahun) dan dewasa (umur 15-59 tahun) dalam periode waktu tertentu yang telah melakukan kegiatan yang berkaitan dengan komputer tertentu (desktop, laptop atau tablet). Data yang digunakan adalah data dalam rentang waktu tahun 2017-2021 yang terdiri dari 34 provinsi. Dapat dilihat pada tabel 1.

Tabel 1. Tabel Proporsi Remaja Dan Dewasa Usia 15-59 Tahun Dengan Keterampilan Teknologi Informasi Dan Komputer (TIK) Menurut Provinsi

Provinsi	Proporsi Remaja Dan Dewasa Usia 15-59 Tahun Dengan Keterampilan TIK Menurut Provinsi (Persen)				
	2017	2018	2019	2020	2021
ACEH	30,56	40,47	46,77	54,25	60,21
SUMATERA UTARA	35,11	43,65	51,78	58,6	67,41
SUMATERA BARAT	38,03	47,49	52,85	58,67	68
RIAU	39,78	49,45	55,37	62,67	70,69
JAMBI	32,8	43,42	50,83	56,87	64,47
SUMATERA SELATAN	32,03	41,33	46,5	54,52	62,59
BENGKULU	32,9	40,42	48,7	53,42	62,1
LAMPUNG	28,36	40,23	48,37	55,57	65,76

Provinsi	Proporsi Remaja Dan Dewasa Usia 15-59 Tahun Dengan Keterampilan TIK Menurut Provinsi (Persen)				
	2017	2018	2019	2020	2021
KEP. BANGKA BELITUNG	35,31	45,45	54,93	60,37	66,33
KEP. RIAU	58,87	65,6	77,18	81,73	89,06
DKI JAKARTA	71,39	77,14	85,17	88,08	91,79
JAWA BARAT	46,09	55,91	65,37	71,09	76,08
JAWA TENGAH	38,75	48,63	58,75	65,78	71,15
DI YOGYAKARTA	57,37	68,82	75,04	81,36	84,72
JAWA TIMUR	38,76	48,07	57,23	63,91	68,07
BANTEN	45,49	57,86	66,96	69,35	75,69
BALI	48,33	57,71	65,48	72,56	77,09
NUSA TENGGARA BARAT	30,04	37,11	47,85	52,72	58,69
NUSA TENGGARA TIMUR	25,3	29,65	36,33	42,89	53,16
KALIMANTAN BARAT	30,38	38,92	47,04	54,1	62,04
KALIMANTAN TENGAH	35,43	43,17	54,54	59,66	66,43
KALIMANTAN SELATAN	37,37	49,32	57,82	62,88	70,39
KALIMANTAN TIMUR	50,56	60,85	69,44	75,33	81,17
KALIMANTAN UTARA	45,68	58,42	65,36	71,99	76,94
SULAWESI UTARA	44,7	51,22	57,48	63,03	69,77
SULAWESI TENGAH	31,7	37,02	44,13	51,68	58,19
SULAWESI SELATAN	38,74	47,07	54,85	60,5	67,29
SULAWESI TENGGARA	35,14	43,94	53,36	60,35	65,75
GORONTALO	34,39	42,71	50,62	55,68	61,94
SULAWESI BARAT	26,24	33,95	40,95	47,66	55,72
MALUKU	31,55	39,2	44,02	49,96	59,26
MALUKU UTARA	25,1	34,24	38,11	45,22	51,53
PAPUA BARAT	34,68	45,41	52,37	59,45	62,31
PAPUA	21,29	24,23	26,45	30,93	30,58

2.2.4 Implementasi *Clustering*

Klasterisasi menggunakan algoritma *K-Means clustering* dimulai dengan melakukah pemilihan jumlah *cluster* k, jumlahnya ditentukan sesuai kebutuhan. Dalam kasus ini *cluster* yang ditentukan adalah 3 yaitu *cluster* tingkat rendah, sedang dan tinggi. Selanjutnya inisialisasi pusat *cluster* atau *centroid*, umumnya dilakukan dengan cara *random*. Pencarian *centroid* awal pada algoritma *K-Means* ini memegang peranan yang penting bagi kualitas *cluster* yang dihasilkan. Pada penelitian ini metode penentuan *centroid* awal yang digunakan adalah metode *Rank Order Centroid (ROC)* yaitu dengan terlebih dahulu mencari *ranking ROC* dari keseluruhan data kemudian menggunakan data *ranking* tertinggi, tengah dan terendah sebagai *centroid* awal. Pada persamaan (1) merupakan rumus pembobotan sebelum menentukan nilai *ROC data*.

$$W_i = \frac{1}{k} \sum_{n=i}^{k} \left(\frac{1}{n}\right) \qquad (1)$$

Keterangan:

W_i = bobot
k = jumlah atribut
i = atribut ke-

Dari rumus pada persamaan (1) untuk pembobotan nilai atribut, didapat perhitungan sebagai berikut:

1) $W_1 = \left(\frac{1}{5}\right) \text{x } (1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5})$
   = 0,457
2) $W_2 = \left(\frac{1}{5}\right) \text{x } (\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5})$
   = 0,257
3) $W_3 = \left(\frac{1}{5}\right) \text{x } (\frac{1}{3} + \frac{1}{4} + \frac{1}{5})$
   = 0,157
4) $W_4 = \left(\frac{1}{5}\right) \text{x } (\frac{1}{4} + \frac{1}{5})$
   = 0,09
5) $W_5 = \left(\frac{1}{5}\right) \text{x } (\frac{1}{5})$
   = 0,04

Selanjutnya didapatkan nilai bobot tiap atribut yang dibulatkan dari perhitungan di atas, tertera pada tabel 2.

Tabel 2. Nilai Bobot Atribut

No.	Atribut (Tahun)	Nilai Bobot Atribut
1.	2017	0,46
2.	2018	0,26
3.	2019	0,16
4.	2020	0,09
5.	2021	0,04

Setelah nilai bobot atribut diperoleh, langkah selanjutnya melakukan perkalian antara nilai bobot atribut dengan data atribut pada *dataset*, maka didapatkan nilai *ROC data* yang

akan menjadi nilai *centroid* awal [7]. Nilai *ROC* tiap data tersebut di dapat dilihat pada tabel 3 berikut.

Tabel 3. Nilai *ROC* per data

Data ke -	Provinsi	ROC Data
1	ACEH	0,183966451
2	SUMATERA UTARA	0,166241586
3	SUMATERA BARAT	0,162688256
4	RIAU	0,155251278
5	JAMBI	0,172194317
6	SUMATERA SELATAN	0,179712602
7	BENGKULU	0,180027938
8	LAMPUNG	0,176317794
9	KEP. BANGKA BELITUNG	0,164185018
10	KEP. RIAU	0,118639179
11	DKI JAKARTA	0,110440408
12	JAWA BARAT	0,139293211
13	JAWA TENGAH	0,152371533
14	DI YOGYAKARTA	0,121745174
15	JAWA TIMUR	0,157286214
16	BANTEN	0,139622818
17	BALI	0,136900418
18	NUSA TENGGARA BARAT	0,18866995
19	NUSA TENGGARA TIMUR	0,223403422
20	KALIMANTAN BARAT	0,182318374
21	KALIMANTAN TENGAH	0,165524504
22	KALIMANTAN SELATAN	0,155223585
23	KALIMANTAN TIMUR	0,130340861
24	KALIMANTAN UTARA	0,137664293
25	SULAWESI UTARA	0,153412166
26	SULAWESI TENGAH	0,191744804
27	SULAWESI SELATAN	0,161390983
28	SULAWESI TENGGARA	0,166172914
29	GORONTALO	0,175980936
30	SULAWESI BARAT	0,206113198
31	MALUKU	0,19065414
32	MALUKU UTARA	0,218760978
33	PAPUA BARAT	0,170968915
34	PAPUA	0,334780317

Penentuan nilai *centroid* awal berdasarkan kebutuhan *cluster* yang datanya meliputi, C1 merupakan data dengan nilai *ROC* terendah, C2 dengan nilai tengah *ROC* dan C3 dengan nilai *ROC* tertinggi. Nilai *centroid* yang didapatkan di antaranya adalah data untuk C1 yaitu provinsi DKI Jakarta dengan nilai ROC 0,110440408, data untuk C2 yaitu provinsi Sulawesi Tenggara dengan nilai ROC 0,166172914 dan data untuk C3 yaitu provinsi Papua dengan nilai ROC 0,334780317.

Setelah mendapatkan *centroid* awal, selanjutnya adalah mencari nilai *Euclidean distance* (D) atau jarak data dari masing-masing data ke nilai *centroid*. Rumus *Euclidean distance* terdapat di persamaan (2).

$$D(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2)$$

Keterangan:

D = selisih jarak data pada titik x dan y
x = nilai data ke-i
y = nilai *centroid* data ke-i
n = jumlah atribut data

Kemudian dilakukan pengelompokan data ke *cluster-cluster* yang dibutuhkan yaitu *cluster* rendah, *cluster* sedang, dan *cluster* tinggi berdasarkan nilai *Euclidean distance*. Jarak data terpendek di tiap data yang akan dikelompokkan ke klusternya, antara C1, C2 ataupun C3. Setelah pengelompokan data, maka keseluruhan proses itu adalah iterasi 1.

Tabel 4. Iterasi 1

Data ke -	D ke C_1	D ke C_2	D ke C_3	Jarak Terpendek	*Cluster*
1	81,413	12,014	46,737	12,014	2
2	70,825	2,8982	57,722	2,8982	2
3	66,839	5,3944	60,729	5,3944	2
4	61,195	9,2583	66,562	9,2583	2
5	74,336	5,0888	54,005	5,0888	2
6	79,409	10,369	48,897	10,369	2
7	79,011	10,024	48,901	10,024	2
8	79,389	10,367	51,296	10,367	2
9	68,092	2,2607	60,023	2,2607	2
10	20,038	50,991	108,16	20,038	1
11	0	69,823	125,58	0	1
12	44,926	25,089	82,562	25,089	2
13	59,157	11,078	69,311	11,078	2
14	21,53	48,831	105,48	21,53	1
15	61,946	7,9495	65,682	7,9495	2
16	44,544	25,8	82,878	25,8	2
17	41,88	28,074	85,274	28,074	2

Data ke -	D ke C_1	D ke C_2	D ke C_3	Jarak Terpendek	*Cluster*
18	83,97	14,529	44,333	14,529	2
19	101,49	32,472	28,213	28,213	3
20	81,463	11,859	47,423	11,859	2
21	69,542	1,7343	58,808	1,7343	2
22	61,419	9,041	66,916	9,041	2
23	34,958	35,277	92,61	34,958	1
24	43,397	26,935	84,339	26,935	2
25	57,125	13,591	69,323	13,591	2
26	85,575	16,65	42,159	16,65	2
27	65,195	5,2317	62,087	5,2317	2
28	69,823	0	58,369	0	2
29	75,425	6,7755	51,897	6,7755	2
30	93,777	24,386	35,23	24,386	2
31	85,037	16,513	42,694	16,513	2
32	98,263	29,302	29,896	29,302	2
33	71,547	3,9995	55,856	3,9995	2
34	125,58	58,369	0	0	3

Jarak terpendek antara data yang menjadi penentuan untuk mengelompokkan data ke *cluster*, seperti yang terlihat di tabel 4. Pada data ke-1 jarak terpendeknya adalah terhadap nilai *centroid* 2, maka data ke-1 termasuk *cluster* 2. Selanjutnya proses iterasi ini diulang kembali dimulai dari penentuan *centroid*. Tidak seperti sebelumnya (*centroid* awal), untuk menentukan *centroid* berikutnya dilakukan dengan menghitung rata-rata dari nilai data di masing-masing *cluster* yang telah dikelompokkan. Proses *K-Means* tersebut dilakukan kembali hingga data yang dikelompokkan tidak berubah dari iterasi sebelumnya, yang berarti perhitungan dinyatakan selesai [14].

Setelah proses iterasi diulang terus-menerus didapatkan nilai *centroid* akhir yang ada di tabel 5. Dari rata-rata nilai *centroid* akhir dapat ditentukan bahwa data di *cluster* 1 menunjukkan *cluster* tingkat tinggi, *cluster* 2 menunjukkan *cluster* tingkat sedang dan *cluster* 3 menunjukkan *cluster* tingkat rendah.

Tabel 5. Nilai *Centroid* Akhir

Centroid	Atribut				
	2017	**2018**	**2019**	**2020**	**2021**
C1	52,9725	62,7888	71,25	76,44	81,568
C2	31,9379	40,1542	47,34	53,11	59,535
C3	24,4825	30,5175	35,46	41,68	47,748

Iterasi berhenti di iterasi keempat yang menunjukkan tidak ada perbedaan *cluster* dengan iterasi sebelumnya. Dari proses iterasi 4 menghasilkan *cluster* 1 terdapat 8 anggota, *cluster* 2 terdapat 22 anggota. dan *cluster* 3 terdapat 4 anggota.

2.2.5 Evaluasi *Clustering*

Evaluasi *clustering* yang akan dilakukan menggunakan metode *Davies-Bouldin Index* terdapat empat tahapan, sebagai berikut.

1) Perhitungan *Sum of Square Within Cluster (SSW)*
   SSW merupakan keterikatan anggota klaster, di mana lebih kecil lebih mirip dan karenanya lebih baik. Perhitungan ini untuk menentukan matriks/kohesi/homogenitas. Kohesi adalah keterikatan anggota klaster dalam satu klaster.

$$SSW_i = \frac{1}{m_i}\sum_{j=i}^{m_i} d(x_j, c_i) \qquad (3)$$

   Keterangan:
   m_i = jumlah data dalam klaster ke-i
   x_j = data pada klaster tersebut
   c_i = *centroid* klaster ke-i
   $d(x_j, c_i)$ = jarak data ke *centroid*
   Hasil perhitungan menggunakan rumus persamaan (3) terlihat di tabel 6.

Tabel 6. Nilai *SSW*

Cluster	*SSW*
1	13,522935
2	11,294253
3	11,988898

2) Menghitung *Sum of Square Between Cluster (SSB)*
   SSB adalah persamaan untuk mengetahui nilai separasi antara klaster. Separasi adalah jarak antara satu klaster dengan klaster lainnya.

$$SSB_{i,j} = d(c_i, c_j) \qquad (4)$$

   Keterangan:
   $d(c_i, c_j)$ = jarak antar *centroid*
   Hasil dari perhitungan nilai *SBB* tertuang di tabel 7 berikut.

Tabel 7. Nilai *SSB*

SSB	*Centroid*		
	1	**2**	**3**
1	0	50,557	74,069
2	50,557	0	20,527
3	74,069	20,527	0

3) Menghitung Rasio
   Setelah didapatkan nilai separasi dan kohesi, kemudian melakukan perhitungan rasio untuk mendapatkan nilai seberapa baik satu klaster dibandingkan dengan klaster lain.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \qquad (5)$$

Keterangan:

$R_{i,j}$ = rasio antar *cluster*
SSW_i = *cluster* 1
SSW_j = *cluster* 2
$SSB_{i,j}$ = separasi dari *cluster* 1 dan 2

Berikutnya melakukan perhitungan nilai rasio dengan rumus persamaan (5). Karena terdapat 3 *cluster*, maka jumlah rasionya juga 3.

Tabel 8. Nilai Rasio

	Rasio
R1,2	0,4908753
R2,3	1,1342905
R1,3	0,3444347

4) Perhitungan *DBI (Davis-Bouldin Index)*

Apabila hasil berdasarkan perhitungan *DBI* yang didapatkan semakin kecil dan atau mendekati nol tetapi tidak negatif (non-negatif >= 0), maka nilai hasil *clustering* semakin baik.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \qquad (6)$$

Keterangan:

k = jumlah *cluster* yang ada
$R_{i,j}$ = rasio antara *cluster* i dan j
max = rasio antar *cluster* terbesar

Hasil nilai *Davies-Bouldin Index (DBI)* yang semakin rendah, maka semakin baik kualitas *cluster* dari *clustering* data menggunakan metode yang dipakai [10][15].

Dari nilai rasio yang tertera di tabel 8. Diambil nilai terbesar rasio untuk menghitung nilai *DBI* menggunakan rumus persamaan (6). Didapatkan nilai *DBI* sebesar 0,163625.

2.2.6 Hasil

Hasil dan kesimpulan dari tahapan penelitian yang telah dilakukan. Berupa pengetahuan yang berkaitan dengan tujuan dari penelitian ini dilakukan.

3. HASIL DAN PEMBAHASAN

Hasil dari proses *K-Means clustering* yang dimulai dari penentuan *centroid* awal menggunakan metode *ROC*. Penentuan nilai *centroid* awal berdasarkan kebutuhan *cluster* yang datanya meliputi, C1 merupakan data dengan nilai *ROC* terendah yaitu provinsi DKI Jakarta dengan nilai ROC 0,110440408, C2 dengan nilai tengah *ROC* yaitu provinsi Sulawesi Tenggara dengan nilai ROC 0,166172914 dan C3 dengan nilai *ROC* tertinggi yaitu provinsi Papua dengan nilai ROC 0,334780317. Data nilai *centroid* awal tertera pada tabel 9.

Tabel 9. Nilai Data *Centroid* Awal

Data ke -	*Centroid* Awal	Atribut				
		2017	2018	2019	2020	2021
11	C1	71,39	77,14	85,17	88,08	91,79
28	C2	35,14	43,94	53,36	60,35	65,75
34	C3	21,29	24,23	26,45	30,93	30,58

Iterasi berhenti di iterasi keempat yang menunjukkan tidak ada perbedaan *cluster* dengan iterasi sebelumnya, hasilnya tertera di tabel 10. Dari proses iterasi 4 dihasilkan bahwa pada *cluster* 1 terdapat 8 anggota, *cluster* 2 terdapat 22 anggota. dan *cluster* 3 terdapat 4 anggota.

Tabel 10. Iterasi 4

Data ke -	D ke C_1	D ke C_2	D ke C_3	Jarak Terpendek	*Cluster*
1	50,479	2,0195	24,028	2,0195	2
2	39,787	11,582	34,997	11,582	2
3	35,995	14,962	38,38	14,962	2
4	30,199	20,693	44,227	20,693	2
5	43,292	7,8791	31,344	7,8791	2
6	48,498	3,663	26,097	3,663	2
7	48,172	3,0856	26,233	3,0856	2
8	48,202	7,6597	28,17	7,6597	2
9	37,003	13,997	37,605	13,997	2
10	12,728	62,882	86,4	12,728	1
11	31,289	81,356	104,59	31,289	1
12	13,71	37,022	60,616	13,71	1
13	27,948	23,321	46,934	23,321	2
14	10,211	60,565	84,007	10,211	1
15	30,892	19,912	43,512	19,912	2
16	13,544	37,508	61,056	13,544	1
17	10,757	39,905	63,437	10,757	1
18	53,091	3,741	21,67	3,741	2
19	70,822	20,511	5,7401	5,7401	3
20	50,463	3,3609	24,457	3,3609	2
21	38,503	12,79	36,283	12,79	2
22	30,231	20,896	44,519	20,896	2
23	3,7732	47,173	70,742	3,7732	1
24	12,171	38,815	62,395	12,171	1
25	26,661	24,317	47,574	24,317	2
26	54,876	4,9028	19,46	4,9028	2

Data ke -	D ke C_1	D ke C_2	D ke C_3	Jarak Terpendek	*Cluster*
27	34,299	16,286	39,869	16,286	2
28	38,759	12,324	35,875	12,324	2
29	44,611	5,9743	29,567	5,9743	2
30	62,933	12,492	12,016	12,016	3
31	54,364	4,6991	19,995	4,6991	2
32	67,579	17,125	6,9379	6,9379	3
33	40,651	10,409	33,76	10,409	2
34	95,75	46,192	23,261	23,261	3

Evaluasi *clustering* menggunakan *DBI* dilakukan mulai dari perhitungan nilai *SSW*, *SSB*, rasio dan terakhir *DBI*. Dari keseluruhan perhitungan tersebut didapatkan nilai *DBI* sebesar 0,163625, yang berarti akurasi dari hasil *clustering* baik karena nilainya mendekati 0. Penelitian oleh [7] juga menghasilkan akurasi yang baik. Maka dari itu, penelitian ini yang menggunakan *dataset* berbeda pun mendapatkan hasil *clustering* yang baik.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan menggunakan metode *K-Means clustering* dengan *Rank Order Centroid (ROC)* untuk *centroid* awal dan pengujian *Davies-Bouldin Index (DBI).* Didapatkan hasil dari *clustering* yang terdiri dari *cluster* 1 merupakan *cluster* dengan tingkat proporsi individu yang memiliki keterampilan TIK tinggi terdapat 8 provinsi yaitu, provinsi Riau, DKI Jakarta, Jawa Barat, DI Yogyakarta, Banten, Bali, Kalimantan Timur dan Kalimantan Utara. *Cluster* 2 merupakan *cluster* dengan tingkat proporsi individu yang memiliki keterampilan TIK sedang memiliki anggota terbanyak di mana terdapat 22 provinsi yaitu provinsi Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep. Bangka Belitung, Jawa Tengah, Jawa Timur, NTB, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Maluku, dan Papua Barat. *Cluster* 3 merupakan *cluster* dengan tingkat proporsi individu yang memiliki keterampilan TIK rendah terdapat 4 provinsi yaitu, provinsi NTT, Sulawesi Barat, Maluku Utara, dan Papua. Dari hasil *clustering* di dapatkan nilai pengujian *DBI* 0,163625, yang berarti kualitas dari hasil *clustering* baik. Maka dari itu, *K-Means* dengan metode *ROC* untuk penentuan *centroid* awal cukup efektif untuk digunakan. Dari hasil *clustering* ini, informasi tersebut dapat menjadi masukan bagi pemerintah mengenai wilayah yang termasuk tingkat rendah perlu dilakukan perkembangan supaya dapat menjaga dan meningkatkan kualitas sumber daya manusia dalam bidang TIK. Sebagai pertimbangan untuk penelitian selanjutnya ada baiknya untuk menggunakan algoritma *clustering* lain sebagai perbandingan dan atau menggunakan metode evaluasi *clustering* lain.

DAFTAR PUSTAKA

[1] D. Wilandini and Purwantoro, "Penerapan Algoritma Naive Bayes dalam Mengklasifikasikan Media Sosial Untuk Mengamati Trend Kuliner," *J. Teknol. Terpadu*, vol. 8, no. 1, pp. 31–39, 2022, doi: https://doi.org/10.54914/jtt.v8i1.535.

[2] A. Voutama, U. Enri, I. Maulana, and E. Novalia, "Sosialisasi Literasi Digital Bagi Remaja dan Calistung Untuk Anak-Anak di Desa Telukbuyung Karawang," *J. Pemberdaya. Komunitas MH Thamrin*, vol. 4, no. 1, pp. 34–41, 2022, doi: 10.37012/jpkmht.v4i1.870.

[3] R. Muntaqo, "Teknologi Informasi dan Komunikasi Dalam Perkembangan Budaya Masyarakat," *J. Penelit. dan Pengabdi. Kpd. Masy. UNSIQ*, vol. 4, no. 1, pp. 12–20, 2017, doi: https://doi.org/10.32699/ppkm.v4i1.401.

[4] A. Purwanto, A. Primajaya, and A. Voutama, "Penerapan Algoritma C4.5 Dalam Prediksi Potensi Tingkat Kasus Pneumonia Di Kabupaten Karawang," *J. Sist. dan Teknol. Inf.*, vol. 8, no. 4, p. 390, 2020, doi: 10.26418/justin.v8i4.41959.

[5] A. Yoga Pratama, Y. Umaidah, and A. Voutama, "Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor Dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja)," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 897–910, 2021, doi: http://dx.doi.org/10.30645/j-sakti.v5i2.386.

[6] F. Yunita, "Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru," *Sistemasi*, vol. 7, no. 3, p. 238, 2018, doi: 10.32520/stmsi.v7i3.388.

[7] P. Sari, "Analisis Kinerja Algoritma K-Means Dengan Penentuan Centroid Menggunakan Metode Rank Order Centroid (Roc)," Universitas Sumatera Utara, 2020.

[8] H. Irwandi, O. S. Sitompul, and S. Sutarman, "K-Means Performance Optimization Using Rank Order Centroid (ROC) And Braycurtis Distance," *SinkrOn*, 2022, doi: 10.33395/sinkron.v7i2.11371.

[9] M. Herviany, S. P. Delima, T. Nurhidayah, and Kasini, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokkan Daerah Rawan Tanah Longsor di Provinsi Jawa Barat," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 34–40, 2021.

[10] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 9, no. 3, pp.

102–109, 2019, doi: 10.31940/matrix.v9i3.1662.

[11] A. Chusyairi and P. Ramadar Noor Saputra, "Pengelompokan Data Puskesmas Banyuwangi Dalam Pemberian Imunisasi Menggunakan Metode K-Means Clustering," *Telematika*, vol. 12, no. 2, pp. 139–148, 2019, doi: 10.35671/telematika.v12i2.848.

[12] V. Ramadhan and A. Voutama, "Clustering Menggunakan Algoritma K-Means Pada Penyakit ISPA di Puskesmas Kabupaten Karawang," *J. Pendidik. dan Konseling*, vol. 4, pp. 462–473, 2022, doi: https://doi.org/10.31004/jpdk.v4i5.6632.

[13] P. Alam Jusia, F. Muhammad Irfan, and Kurniabudi, "Clustering Data Untuk Rekomendasi Penentuan Jurusan Perguruan Tinggi Menggunakan Metode K-Means," *J. IKRA-ITH Inform.*, vol. 3, no. 3, p. 75, 2019.

[14] Carudin, "Pemanfaatan DataTransaksi Untuk Dasar Membangun Strategi Berdasarkan Karakteristik Pelanggan Dengan Algoritma K-Means Clustering dan Model RFM," *J. Teknol. Terpadu*, vol. 7, no. 1, pp. 15–22, 2021, doi: https://doi.org/10.54914/jtt.v7i1.318.

[15] I. W. Septiani, A. C. Fauzan, and M. M. Huda, "Implementasi Algoritma K-Medoids Dengan Evaluasi Davies-Bouldin- Index Untuk Klasterisasi Harapan Hidup Pasca Operasi Pada Pasien Penderita Kanker Paru-Paru," *J. Sist. Komput. dan Inform.*, vol. 3, no. 4, pp. 556–566, 2022, doi: 10.30865/json.v3i4.4055.