

HASIL CEK_Kurnia, Zahrotun, Linarti_K-Medoids, Silhouette Coefficient, One Hot Encoding, Euclidean Distance, Cluster

by Herri Kurnia, Lisna Zahrotun, Utaminingsih Linarti Pengelompokan
Mahasiswa Berdasarkan Data Akademik

Submission date: 20-Sep-2022 12:41PM (UTC+0700)

Submission ID: 1904318680

File name: sarkan_Data_Akademik_Sebelum_Kuliah_dan_Masa_denga_K-Medoids.pdf (932.69K)

Word count: 4180

Character count: 23650



Available online at : <http://bit.ly/InfoTekJar>
**InfoTekJar : Jurnal Nasional Informatika dan
 Teknologi Jaringan**

ISSN (Print) 2540-7597 | ISSN (Online) 2540-7600



Pengelompokan Mahasiswa Berdasarkan Data Akademik Sebelum Kuliah dan Masa Studi Menggunakan *K-Medoids*

Herri Kurnia, Lisna Zahrotun, Utaminingsih Linarti

Universitas Ahmad Dahlan Fakultas Teknologi Industri Program Studi Teknik Informatika, Jl. Ringroad Selatan, Yogyakarta, 55191., Indonesia

KEYWORDS

K-Medoids, Silhouette Coefficient, One Hot Encoding, Euclidean Distance, Cluster

CORRESPONDENCE

Phone: +62 81328265007

E-mail: lisna.zahrotun@tif.uad.ac.id

ABSTRACT

The research objective is to obtain information about the results of grouping that are useful for the campus, especially study programs, to be used as consideration for future admissions by grouping, causing a mismatch between the number of students and the existing campus facilities. This research was conducted at university X which has several faculties, one of which is faculty Y which consists of the R study program, S study program, T study program and U study program. This research uses the K-Medoids method. The stages of this research started with load dataset, data cleaning, data selection, data transformation with one hot encoding, euclidean distance, and k-medoids to produce clusters. Testing the quality of the clusters in this study using the silhouette coefficient. The research resulted in recommended student data and all of them came from Java Island. In the dataset of study programs R, S, and U, the recommended data are obtained with a total number of 9, 57, and 64, respectively, which have an average math score of at least 82. Meanwhile, for the T study program dataset, there are 35 data with an average mathematical value. amounted to 73.89. The test results for the dataset of study programs R, S, T and U are 0.52, 0.67, 0.35, and 0.65 respectively, so the results are quite good.

ABSTRAK

Tujuan penelitian yaitu untuk memperoleh informasi tentang hasil pengelompokan yang berguna bagi pihak kampus terutama prodi untuk dijadikan bahan pertimbangan untuk penerimaan mahasiswa baru kedepannya dengan cara mengelompokkan, sehingga menyebabkan ketidaksesuaian antara jumlah mahasiswa dengan fasilitas kampus yang ada. Penelitian ini dilakukan pada universitas X memiliki beberapa fakultas salah satunya fakultas Y yang terdiri dari program studi R, program studi S, program studi T dan program studi U. Penelitian ini menggunakan metode *K-Medoids*. Tahapan penelitian ini dimulai dari *load* dataset, data *cleaning*, data *selection*, transformasi data dengan *one hot encoding*, *euclidean distance*, dan *k-medoids* untuk menghasilkan *cluster*. Pengujian kualitas *cluster* dalam penelitian ini menggunakan *silhouette coefficient*. Penelitian menghasilkan data mahasiswa yang direkomendasikan dan semuanya berasal dari Pulau Jawa. Pada dataset program studi R, S, dan U diperoleh data yang direkomendasikan dengan jumlah berturut-turut 9, 57, dan 64 yang memiliki rata-rata nilai matematika minimal 82. Sedangkan untuk dataset program studi T terdapat 35 data dengan rata-rata nilai matematika sebesar 73,89. Hasil pengujian untuk dataset program studi R, S, T dan U berurut-turut sebesar 0,52, 0,67, 0,35, dan 0,65 sehingga dinyatakan hasilnya cukup bagus.

PENDAHULUAN

Universitas X mempunyai beberapa fakultas, salah satunya Fakultas Y. Fakultas Y memiliki 4 program studi yaitu program studi R, program studi S, program studi T dan program studi U. Peningkatan jumlah mahasiswa baru terjadi setiap tahun untuk semua program studi di Fakultas Y. Namun peningkatan tersebut tidak diimbangi dengan prosentase jumlah kelulusan mahasiswa tepat waktu di setiap program studi, yang secara agregat menjadi prosentase kelulusan mahasiswa di Fakultas Y. Salah satu faktor yang menyebabkan rendahnya prosentase kelulusan mahasiswa tepat waktu antara lain: (1) kurang minat pada pilihan prodi, (2) masih mengikuti kuliah, (3) metode bimbingan yang tidak intensif, (4) kurang lengkapnya fasilitas kampus dan (5) pengaruh lingkungan [1].

Tabel 1. Data Jumlah Mahasiswa dan Kelulusan Fakultas Y

Tahun	Jumlah Mahasiswa	Jumlah Lulusan
2012	360	14 atau 4%
2013	340	22 atau 6%
2014	591	55 atau 9%
2015	837	181 atau 22%

Adanya ketidakseimbangan tersebut akan berimplikasi pada faktor lain dalam pembelajaran akademik, misalnya (1) tidak seimbang jumlah dosen dan mahasiswa, (2) meningkatnya penggunaan fasilitas kampus, seperti laboratorium yang tidak sesuai dengan jumlah mahasiswa sehingga membutuhkan penambahan fasilitas atau penambahan jam slot praktikum. Selama ini data mahasiswa sebelum masuk seperti nilai raport, asal sekolah dan asal kabupaten dengan data kelulusan mahasiswa yaitu lama studi, TOEFL dan IPK belum pernah dikelompokkan berdasarkan aturan tertentu, hal ini dijadikan bahan pertimbangan bagi pihak Fakultas Y khususnya setiap program studi dalam menyeleksi mahasiswa baru.

Data yang digunakan merupakan data yang cukup besar karena mencakup semua mahasiswa angkatan 2014 – 2015 dari program studi R, S, T dan U. *Data Mining* merupakan suatu teknik yang diperuntukan untuk memproses penggalian informasi yang belum diketahui dalam suatu dataset atau basis data yang besar [2]. *Data Mining* mempunyai banyak teknik pengolahan dalam mengolah data, salah satu tekniknya yaitu Teknik *Clustering*.

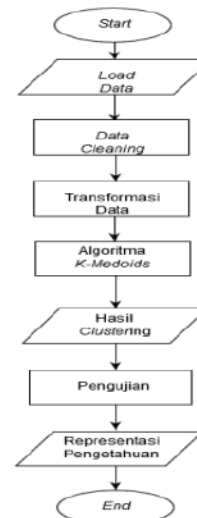
Clustering adalah sebuah teknik untuk memproses data dan mengelompokkan data dalam suatu kelas yang memiliki kemiripan objek [3]. Pada Teknik *Clustering* terdapat beberapa algoritma yang dapat mengelompokkan data dengan baik, diantaranya K-Means [4] [5], Shared Nearest Neighbour (SNN) [6] dan Algoritma *K-Medoids*. Penelitian ini menggunakan Metode *K-Medoids* dengan pendekatan *Euclidean Distance*. Algoritma *K-Medoids* efektif untuk mengatasi dataset yang mana ada beberapa data (outlier) yang terletak sangat jauh dari kebanyakan data lainnya [7]. Pengolahan data pada penelitian ini menggunakan Algoritma *K-Medoids* dengan pendekatan *Euclidean Distance* dengan data yang digunakan bertipe numerik dan teks.

Berdasarkan permasalahan yang ada, disimpulkan bahwa untuk mengatasi masalah mahasiswa yang belum lulus tepat waktu maka diperlukan pengelompokkan menggunakan Algoritma *K-Medoids* berdasarkan ketentuan – ketentuan tertentu. Sehingga hasil

pengelompokkan tersebut dapat memberikan informasi dan juga menjadi bahan pertimbangan dalam penyeleksian calon mahasiswa baru yang diharapkan membantu Fakultas Y khususnya setiap program studi yang ada di fakultas tersebut.

METODE

Metode yang digunakan dalam penelitian ini adalah metode *K-Medoids*. *K-Medoids* adalah algoritma *clustering* yang menggunakan objek sebagai pusat *cluster* pada setiap *cluster*. Algoritma *K-Medoids* menggunakan objek sebagai perwakilan (medoid) sebagai pusat cluster untuk setiap cluster, sedangkan K-Means menggunakan nilai rata-rata (mean) sebagai pusat cluster [8]. Tahapan pada penelitian ini melalui beberapa tahap yang dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Penjelasan dari tahap-tahap dalam penelitian ini, sebagai berikut

1. **Load Data**
Load Data merupakan proses mengambil data dari dataset yang sudah ditentukan berdasarkan kebutuhan sedemikian rupa
2. **Data Cleaning**
Data Cleaning adalah sebuah proses pembersihan data dengan cara memilih data atau menghapus data yang sesuai dengan ketentuan yang bertujuan untuk menghindari data yang *noise*.
3. **Transformasi Data**
Transformasi Data merupakan teknik untuk mengubah data menjadi bentuk lain yang diasumsikan memenuhi analisis ragam. Teknik transformasi yang digunakan pada penelitian ini menggunakan Teknik *One Hot Encoding*. *One Hot Encoding* adalah sebuah cara untuk mentransformasikan atau merepresentasikan data agar dapat dipahami komputer. Prinsip kerja metode ini yaitu dengan membuat sebuah array 1 dimensi yang memiliki panjang sebanyak jenis class yang ada dan memiliki nilai biner [9].
4. **Algoritma *K-Medoids***
Data akan diambil secara acak untuk dijadikan data pusat pada *cluster*, setiap data berpeluang menjadi data pusat tetapi data yang paling tengahlah yang dijadikan data pusat pada suatu *cluster*

berdasarkan ketentuan dari Algoritma *K-Medoids*. Langkah – langkah Algoritma *K-Medoids*, sebagai berikut :

- a. Inisialisasi pusat cluster sebanyak k (jumlah cluster).
- b. Kelompokkan setiap data ke cluster terdekat menggunakan pendekatan *Euclidian Distance* untuk menghitung jarak antar data dengan persamaan:

$$d(x, y) = | |x - y| | = \sqrt{(\sum_{i=1}^n (x_{(i)} - y_{(i)}))^2}; 1, 2, 3, \dots, n \dots (1)$$

Keterangan :

$x_{(i)}$ = data ke i pertama.

$y_{(i)}$ = data ke i kedua.

n = banyak data.

c. Kemudian pilih data secara random pada setiap cluster yang dijadikan calon medoid baru.

d. Setelah itu, hitung jarak setiap data yang berada pada masing-masing cluster dengan calon medoid baru.

e. Kemudian menghitung total simpangan (S) dengan menghitung nilai total distance baru – total distance lama. Jika $S < 0$, maka ganti objek dengan data cluster untuk membentuk sekumpulan k objek baru sebagai *medoid*.

f. Ulangi langkah 3 sampai 5 sampai tidak terjadi perubahan *medoid*, sehingga didapatkan cluster beserta anggota cluster masing-masing.

5. Hasil *Clustering*

Hasil *Clustering* merupakan hasil pengolahan data dari tahap sebelumnya yaitu tahap Algoritma *K-Medoids*, dimana hasil ini berupa data-data yang sudah dikelompokkan berdasarkan tingkat kemiripan antar data.

6. Pengujian

Pada penelitian ini, pengujian menggunakan Metode *Silhouette Coefficient*, yang mana metode ini akan menghitung tingkat kedekatan yang terdapat antar data atau objek dalam suatu *cluster*. Langkah – langkah pada proses *silhouette coefficient* [9], sebagai berikut :

- a. Hitung rata-rata jarak dari suatu dokumen misalkan i dengan semua dokumen lain yang berada dalam satu cluster [10].

$$a(i) = \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i, j) \dots\dots\dots (2)$$

Yang mana j merupakan dokumen lain dalam cluster A dan d(i,j) bermakna jarak antar dokumen i dengan j.

- b. Hitung rata-rata jarak dari dokumen i tersebut dengan semua dokumen di cluster lain, dan diambil nilai terkecilnya [10].

$$d(i, C) = \frac{1}{|A|} \sum_{j \in C} d(i, j) \dots\dots\dots (3)$$

dengan d(i,C) adalah jarak rata-rata dokumen i dengan semua objek pada cluster lain C dimana $A \neq C$ [10].

$$b(i) = \min_{C \neq A} d(i, C) \dots\dots\dots (4)$$

- c. Nilai *Silhouette Coefficient* [11].

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \dots\dots\dots (5)$$

Keterangan :

$s(i)$ = Nilai *Silhouette Coefficient*.

$a(i)$ = Rata-rata jarak i terhadap semua objek di klaster A.

$b(i)$ = Rata-rata jarak i terhadap semua objek klaster lain.

7. Representasi Pengetahuan

Tahap terakhir ini merupakan tahap pengambilan kesimpulan berdasarkan dari hasil pengujian. Hasil pengujian akan berada pada kisaran -1 sampai 1. Jika nilai dari hasil pengujian mendekati nilai 1, maka suatu cluster yang telah dilakukan pengelompokkan data dapat dikatakan baik. Sedangkan jika nilai dari hasil pengujian mendekati nilai -1, maka dapat disimpulkan bahwa *cluster* tersebut buruk dalam mengelompokkan data [12].

HASIL

Penelitian ini menggunakan data mahasiswa dari program studi R, S, U dan T dengan jumlah dari masing-masing data berturut-turut yaitu 31, 90, 90, dan 76. Data tersebut merupakan data mahasiswa angkatan 2014 – 2015 dengan beberapa atribut yaitu Angkatan, NIM, Nama, Prodi, Jalur Masuk, Sekolah, Nama Sekolah, Asal Kabupaten, Wilayah 1, Wilayah 2, Wilayah 3, nilai raport Matematika, Lama Studi, IPK dan TOEFL. Tahapan pemrosesan data diawali dengan dengan tahap load data. Dataset Program Studi S di load terlebih dahulu sebelum diproses seperti pada Tabel Load Data.

Tabel 2. Load Data

Angkatan	NIM	Nama	Prodi	Jalur_Masuk	Sekolah	Nama_Sekolah	asal_kabupaten	Wilayah 1	Wilayah 2	Wilayah 3	MTK	LABA
0	2014	1400019002	MEKAWAMAD ANDRIANI PRATAMA	TS00W IK010194	PMCK-Raport	SMA	SMA Negeri 2 Yogyakarta	DI Yogyakarta	0	1	0	42,00000
1	2014	1400019008	DEDI MUSTAA	TS00W IK010194	PMCK-Raport	SMA	SMA KANANDA DIKARSA 1 KOTABUKA	Lampung	0	1	0	56,00000
2	2014	1400019012	NOVIANUSARA	TS00W IK010194	PMCK-Raport	SMA	SMA Negeri 12 batang	Jawa	0	1	0	67,66667
3	2014	1400019014	MAGLIS SAUDIRI PRHATMOKO	TS00W IK010194	PMCK-Raport	SMA	SMA Negeri 3 Yogyakarta	DI Yogyakarta	0	1	0	80,33333
4	2014	1400019017	MEKAWAMAD HENDRA PUTRA	TS00W IK010194	PMCK-Raport	SMA	sma perse 2 kander lampung	Lampung	0	1	0	66,33333
71	2015	1500019163	IRMADE LUTRI RAMACHARI	TS00W IK010194	PMCK-Raport	SMA	SMA NEGERI 9 TEBO	Jawa	0	1	0	88,33333
72	2015	1500019165	BITANI PRATIWI	TS00W IK010194	PMCK-Raport	SMA	SMA Negeri 1 Sambutan	Jawa Tengah	0	1	0	84,66667
73	2016	1600019188	RAMA YLEH ELIUSIARA	TS00W IK010194	PMCK-Raport	SMA	SMA Busti Utara, Tegal	Jawa Timur	0	1	0	84,00000
74	2015	1500019206	SARA LURA RANJO ELIUSIARA	TS00W IK010194	PMCK-Raport	SMA	SMA Negeri 1 Semburan, Tegal	Bali	0	1	0	85,00000
75	2015	1500019207	DEA ANDANI MELIA	TS00W IK010194	PMCK-Raport	SMA	SMA NEGERI 3 GUNDUH	Jawa Barat	0	1	0	87,00000

Tahap selanjutnya yaitu melakukan proses data *cleaning* tetapi untuk atribut NIM dan Nama Sekolah disimpan. selanjutnya transformasi data menggunakan metode *One Hot Encoding* pada atribut Sekolah. Pada atribut asal kabupaten, data ditransformaasi menjadi 3 yaitu asal 1 yang berwarna abu-abu terdiri Maluku Utara dan Kalimantan Tengah, asal 2 yang berwarna coklat terdiri dari Pulau Jawa, Pulau Sumatra, Pulau Sulawesi, Kalimantan Barat, Kalimantan Selatan, Bali, NTT dan NTB, asal 3 yang berwarna hijau terdiri dari Pulau Papua, Kalimantan Timur dan Maluku. Pengelompokkan ini berdasarkan tingkat kualitas pendidikan dengan parameter tertentu.



Gambar 2. Pemetaan Kualitas Pendidikan di Indonesia[13]

Atribut asal sekolah diubah menjadi 3 atribut baru yaitu wilayah I, wilayah II, dan wilayah III yang disisipkan kedalam dataset beserta nilai yang sudah diperoleh dari hasil transformasi berdasarkan pemetaan kualitas pendidikan di Indonesia.

Tabel 3. Hasil Pemrosesan Data *Cleaning &* Transformasi Data

Wilayah I	Wilayah II	Wilayah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK
0	0	1	0.42550000	1748	3.42	466	0	0	1
1	0	1	0.56000000	1749	2.91	470	0	0	1
2	0	1	0.87666667	1513	3.09	413	0	1	0
3	0	1	0.80333333	1513	3.34	410	0	0	1
4	0	1	0.86333333	1642	3.22	456	0	1	0
...
71	0	1	0.88333333	1455	3.48	406	0	1	0
72	0	1	0.84666667	1455	3.66	413	0	1	0
73	0	1	0.84000000	1478	3.30	463	0	1	0
74	0	1	0.85000000	1455	3.44	406	0	1	0
75	0	1	0.87000000	1455	3.57	436	0	1	0

76 rows x 10 columns

Setelah didapatkan hasil pemrosesan tersebut, selanjutnya masuk ketahap Algoritma *K-Medoids*, dimana setiap data dihitung jarak kedekatannya dengan data lain dengan menggunakan metode *Euclidean Distance*.

Data ke -1,2 :

$$d(1,2) = \sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (42.55 - 56.0)^2 + (1748 - 1749)^2 + (3.42 - 2.91)^2 + (466 - 470)^2} = 14.077$$

⋮
⋮
⋮
⋮
⋮

Data ke - 1, 76 :

$$d(1,76) = \sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (42.55 - 87.0)^2 + (1748 - 1455)^2 + (3.42 - 3.57)^2 + (466 - 436)^2} = 297.87$$

Hasil dari perhitungan jarak setiap data / *Euclidean Distance* tersebut yang akan menjadi dataset awal untuk melakukan proses *K-Medoids*.

Tabel 4. Hasil Perhitungan *Euclidean Distance*

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14
31	0.00000	14.077024	245.000000	244.917040	104.874200	239.707141	243.430002	164.688915	60.236912	142.160079	...	285.955748	295.45869	
32	14.077024	0.000000	244.846500	244.720950	98.029182	238.870459	243.407121	164.673766	54.991950	141.659704	...	285.908604	294.96701	
33	245.000000	244.846500	0.000000	8.952546	101.806636	37.062304	6.708360	90.227957	279.517145	116.800264	...	42.014166	83.80299	
34	244.917040	244.720950	8.952546	0.000000	102.257450	40.876680	6.910153	60.911380	278.640568	116.488930	...	42.010172	83.84425	
35	104.874200	98.029182	101.806636	102.257450	0.000000	539.000382	391.400004	246.516962	50.208648	329.226000	...	373.483793	387.42901	
...	
072	300.567542	300.621486	58.425996	58.702807	390.221616	72.815772	59.220099	147.328657	336.91713	174.172022	...	17319615	67.11630	
073	300.722248	300.847703	58.000031	58.257010	389.365376	68.837766	58.107290	147.632250	336.170984	174.484028	...	16.057562	60.15170	
074	270.188160	270.536515	61.142181	60.950259	364.074767	37.466727	58.648665	158.133797	312.900917	163.621902	...	50.027319	25.98704	
075	300.001100	300.203446	58.482762	58.341990	390.218940	72.836896	59.800636	147.182714	336.810023	174.114025	...	17.482991	67.16620	
076	297.670444	297.582853	62.398919	62.825711	387.217183	59.666162	61.044321	151.484188	325.172052	177.722025	...	28.038207	37.05672	

76 rows x 14 columns

Data hasil perhitungan *Euclidean Distance* diproses dengan algoritma *K-Medoids* dan menghasilkan beberapa *cluster* beserta anggota – anggota yang mirip dengan *medoid* atau data pusatnya yang menjadi hasil akhir dari clustering. Data yang disimpan diawal tadi yaitu data NIM dan Nama Sekolah dikembalikan sesuai masing – masing data.

Tabel 5. Hasil *Clustering Program Studi S*

NIM	Nama_Sekolah	Wilayah I	Wilayah II	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	Cluster	
0	1400019802	SMK Negeri 2 Yogyakarta	0	1	0	42.550000	1748	3.42	466	0.0	0.0	1.0
1	1400019000	SMK NEWIYA DHARMA 1 KOTAMULIA	0	1	0	56.000000	1749	2.91	470	0.0	0.0	1.0
2	1400019010	SMK Negeri 13 Pekanbaru	0	1	0	87.666667	1513	3.09	413	0.0	1.0	0.0
3	1400019014	SMK Negeri 1 Yogyakarta	0	1	0	80.333333	1513	3.34	410	0.0	0.0	1.0
4	1400019017	sma perintis 2 bandar lampung	0	1	0	86.333333	1642	3.22	456	0.0	1.0	0.0
...
71	1500019163	SMA NEGERI 5 TEBU	0	1	0	88.333333	1455	3.48	406	0.0	1.0	0.0
72	1500019165	SMA Negeri 1 Bandungan	0	1	0	84.666667	1455	3.66	413	0.0	1.0	0.0
73	1500019166	SMK Budi Utomo Perak	0	1	0	84.000000	1478	3.30	463	0.0	1.0	0.0
74	1500019205	SMA Negeri 1 Tombulahu Huku	0	1	0	85.000000	1455	3.44	406	0.0	1.0	0.0
75	1500019207	SMA NEGERI 2 CREBON	0	1	0	87.000000	1455	3.57	436	0.0	1.0	0.0

76 rows x 13 columns

Setiap Dataset melalui proses yang sama seperti paparan diatas dan menghasilkan hasil akhir yang berupa data-data yang sudah di-clustering.

Tabel 6. Hasil *Clustering* Program Studi R

NIM	Nama_Sekolah	Wilyah I	Wilyah II	Wilyah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	Cluster	
0	1400022003	SMN Negeri 1 Bitan	0	1	0	53,67	1451	3,41	408	0	0	1	2
1	1400022005	SMK Negeri 2, Pekanbaru	0	1	0	81,00	1395	3,32	406	0	0	1	1
2	1400022010	SMA Negeri 2, Kallanda	0	1	0	79,67	1679	3,19	470	0	1	0	0
3	1400022024	SMK Hasanah Pekanbaru	0	1	0	78,67	1729	3,34	416	0	0	1	0
4	1400022032	sma negeri 1 bandar sribaharu	0	1	0	78,67	1513	3,44	420	0	1	0	2
5	1400022039	SMA NEGERI 1 CIAWIGEBANG	0	1	0	78,67	1451	3,79	440	0	1	0	2
6	1400022042	SMA Negeri 1, Marabahan	0	1	0	84,67	1455	3,65	426	0	1	0	2
7	1400022048	SMA N2 FANFAK	0	0	1	83,33	1749	3,02	426	0	1	0	0
8	1400022048	SMA Negeri 1, Cibarah	0	1	0	77,33	1513	3,21	426	0	1	0	2
9	1400022056	smi muh 1 bambangluro	0	1	0	81,33	1513	3,25	426	0	0	1	2
10	1400022058	SMK Negeri 2, Yogyakarta	0	1	0	83,33	1382	3,74	420	0	0	1	1
11	1400022079	SMA N 1 TEMPEL	0	1	0	83,33	1834	3,31	443	0	1	0	0
12	1400022080	smi muhammadiah 2 martoyudan	0	1	0	83,67	1749	3,23	416	0	0	1	0
13	1500022093	SMA Negeri 2, Kelasang	0	1	0	78,67	1398	3,85	473	0	1	0	1
14	1500022098	SMA Negeri 1, Hekak	0	1	0	78,67	1455	3,02	450	0	1	0	2
15	1500022010	SMA Negeri 1, Tenale	1	0	0	83,00	1471	3,25	456	0	1	0	2
16	1500022019	SMA Muhammadiyah, Ilorontalo	0	1	0	82,67	1494	3,02	410	0	1	0	1
17	1500022027	SMA Negeri 3, Slawi	0	1	0	78,67	1439	3,82	408	0	1	0	2
18	1500022031	MA Negeri, Katasang	0	1	0	77,67	1455	3,83	413	1	0	0	2
19	1500022033	SMK Negeri 1 Sambang	0	1	0	89,33	1371	3,57	420	0	0	1	1
20	1500022034	SMK N 1 BLORA	0	1	0	84,67	1371	3,69	400	0	0	1	1
21	1500022035	SMA Negeri 2, Bantipapan	0	0	1	77,67	1432	3,82	420	0	1	0	2
22	1500022037	SMA Negeri 1, Sengah Tenila	0	1	0	79,33	1455	3,87	446	0	1	0	2
23	1500022042	SMA NEGERI 12 MERAANGIN	0	1	0	84,67	1471	3,48	446	0	1	0	2
24	1500022047	SMK Negeri 1, Setayu	0	1	0	81,67	1459	3,35	403	0	0	1	2
25	1500022053	SMA Negeri 2, Dumai	0	1	0	83,33	1455	3,54	416	0	1	0	2
26	1500022058	SMA Negeri 1, Kebumen	0	1	0	83,00	1494	3,65	403	0	1	0	1
27	1500022059	SMA Negeri 1, Mania	0	1	0	85,00	1439	3,51	440	0	1	0	2
28	1500022067	SMK NEGERI 1 TOJONG	0	1	0	81,33	1494	3,59	443	0	0	1	1
29	1500022069	SMAN 4 MUARO JAMBI	0	1	0	88,33	1494	3,41	413	0	1	0	1
30	1500022083	SMK Negeri 2, Pundoloto	0	1	0	78,67	1439	3,65	406	0	0	1	2

Tabel 7. Hasil *Clustering* Program Studi T

NIM	Nama_Sekolah	Wilyah I	Wilyah II	Wilyah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	Cluster	
0	1400022002	SMA Negeri 5, Marna	0	1	0	54,666667	1398	3,60	456	0	1	0	0
1	1400022011	SMAN 1 MONGGO	0	1	0	84,666667	1749	3,52	400	0	1	0	1
2	1400022016	STM Pendidikan, Yogyakarta	0	1	0	58,333333	1398	3,60	456	0	0	1	0
3	1400022017	MA Negeri, Praya	0	1	0	56,333333	1398	3,60	456	1	0	0	0
4	1400022018	SMAN N1 WITA POKDA	0	1	0	55,666667	1398	3,60	456	0	1	0	0
...
85	1500022149	SMA Muhammadiyah 1, Yogyakarta	0	1	0	81,666667	1377	3,51	423	0	1	0	3
86	1500022154	MA Negeri 2, Banjarnegara	0	1	0	62,000000	1377	3,67	423	1	0	0	3
87	1500022165	SMA Negeri 1, Davauluhur	0	1	0	58,000000	1355	3,64	463	0	1	0	0
88	1500022179	SMA Albro, Albro	0	1	0	91,333333	1355	3,74	423	0	1	0	3
89	1500022183	SMA Negeri 2, Muara Bungo	0	1	0	78,500000	1455	3,39	470	0	1	0	1

90 rows x 13 columns

Tabel 8. Hasil *Clustering* Program Studi U

NIM	Nama_Sekolah	Wilyah I	Wilyah II	Wilyah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	Cluster	
0	1400018012	SMA Drajat Mada, Yogyakarta	0	1	0	83,333333	1455	3,46	456	0	1	0	1
1	1400018016	MA Negeri 3, Yogyakarta	0	1	0	77,666667	1455	3,25	400	1	0	0	1
2	1400018017	SMA Negeri 1, Balaikan	0	1	0	89,333333	1513	3,57	406	0	1	0	1
3	1400018026	SMK TARUNA BANUSA DAMIS	0	1	0	73,333333	1830	2,68	403	0	0	1	1
4	1400018041	SMAN 01 WITAPONDA	0	1	0	79,000000	1513	3,63	400	0	1	0	1
...
85	1500018242	MA Negeri 1, Metro	0	1	0	79,000000	1455	3,75	423	1	0	0	1
86	1500018248	SMA Negeri 1, Sungsopenuh	0	1	0	90,666667	1306	3,76	403	0	1	0	1
87	1500018248	SMA Negeri 3, Bedan	0	1	0	81,666667	1455	3,62	440	0	1	0	1
88	1500018255	SMA Negeri 1, Purbalangga	0	1	0	75,666667	1343	3,59	433	0	1	0	1
89	1500018210	SMAN KH Mustika, Sukamanah	0	1	0	82,000000	1439	2,99	416	0	1	0	1

90 rows x 13 columns

Semua hasil *Clustering* setiap program studi dilakukan analisis evaluasi pola dan representasi pengetahuan sehingga dapat ditarik kesimpulan yang menghasilkan informasi yang bermanfaat untuk pihak program studi. Berdasarkan hasil *clustering* dari setiap dataset, maka dapat dilakukan evaluasi pola sebagai berikut :

1. Pada dataset Program Studi R dengan jumlah 31 data diperoleh hasil pengujian yang disajikan pada Tabel 9.

Tabel 9. Hasil Pengujian Dataset Program Studi R

Nilai K (Jumlah Cluster)	Hasil Silhouette Coefficient
2	0,518
3	0,52
4	0,435

Berdasarkan hasil pengujian pada dataset Program Studi R dapat ditarik kesimpulan bahwa pada dataset Program Studi R jumlah *cluster* terbaik yaitu 3. Sehingga diperoleh 3 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut :

- Cluster* 0 memiliki jumlah data sebanyak 5 data dan mayoritas berasal dari SMA, Asal 2, luar pulau jawa serta memiliki rata-rata nilai matematika sebesar 81,5 , TOEFL sebesar 434,2, IPK sebesar 3,21 dengan lama studi lebih dari 4 tahun yaitu 4 tahun 9 bulan.
- Cluster* 1 memiliki jumlah data sebanyak 9 data dan berasal dari Asal 2 tetapi dalam pulau jawa dengan mayoritas berasal dari SMK, serta memiliki rata-rata nilai matematika sebesar 82,92 , TOEFL sebesar 420,88 , IPK sebesar 3,547 dengan lama studi kurang dari 4 tahun yaitu 3 tahun 10 bulan.
- Cluster* 2 memiliki jumlah data sebanyak 17 data dan mayoritas berasal dari SMA, Asal 2, luar pulau jawa serta memiliki rata-rata nilai matematika sebesar 78,62 , TOEFL sebesar 426,23 , IPK sebesar 3,541 dengan lama studi 4 tahun 2 hari.

Hasil dari evaluasi pola pada dataset Program Studi R dengan 3 jumlah *cluster*, maka diperoleh *cluster* 1 sebagai *cluster* yang terbaik untuk rekomendasi tepat waktu jika nilai rata-rata matematika besar sama dengan 82,92 dengan asal dari dalam pulau jawa dan rata – rata IPK sebesar 3,547.

2. Pada dataset Program Studi S dengan jumlah 76 data diperoleh hasil pengujian yang disajikan pada Tabel 10.

Tabel 10. Hasil Pengujian Dataset Program Studi S

Nilai K (Jumlah Cluster)	Hasil Silhouette Coefficient
2	0,67
3	0,36
4	0,27

Berdasarkan hasil pengujian pada dataset Program Studi S dapat ditarik kesimpulan bahwa jumlah *cluster* terbaik pada dataset Program Studi S yaitu 2. Sehingga diperoleh 2 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut:

- Cluster* 0 memiliki jumlah data sebanyak 57 data dan mayoritas berasal dari SMA, Asal 2 tetapi dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 82,84, TOEFL sebesar 425,16, IPK sebesar 3,47 dengan lama studi 4 tahun lebih 8 hari.
- Cluster* 1 memiliki jumlah data sebanyak 19 data dan mayoritas berasal dari SMA, Asal 2 tetapi dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 77,15, TOEFL sebesar 438,89, IPK sebesar 3,31 dengan lama studi 4 tahun 8 bulan.

Hasil dari evaluasi pola pada dataset Program Studi S yang memiliki 2 *cluster*, maka diperoleh *cluster* 0 sebagai *cluster* yang terbaik untuk rekomendasi tepat waktu jika nilai rata-rata matematika besar sama dengan 82,84 dengan asal dari dalam pulau jawa dan rata-rata IPK sebesar 3,47.

3. Pada dataset Program Studi T dengan jumlah 90 data diperoleh hasil pengujian yang disajikan pada Tabel 11.

Tabel 11. Hasil Pengujian Dataset Program Studi T

Nilai K (Jumlah Cluster)	Hasil Silhouette Coefficient
2	0,21
3	0,28
4	0,35

Berdasarkan hasil pengujian pada dataset Program Studi T dapat ditarik kesimpulan bahwa jumlah *cluster* terbaik adalah 4. Sehingga diperoleh 4 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut:

- Cluster* 0 memiliki jumlah data sebanyak 35 data dan mayoritas berasal dari SMA, Asal 2 tetapi dalam pulau jawa serta memiliki rata-rata nilai matematika sebesar 73,89, TOEFL sebesar 456,51, IPK sebesar 3,62 dengan lama studi 3 tahun 8 bulan.
- Cluster* 1 memiliki jumlah data sebanyak 16 data dan mayoritas dari SMA yang berasal dari Asal 2, baik luar pulau jawa maupun dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 66,30, TOEFL sebesar 431,69, IPK sebesar 3,37 dengan lama studi 4 tahun 2 bulan.
- Cluster* 2 memiliki jumlah data sebanyak 12 data dan mayoritas dari SMA yang berasal dari Asal 2, baik luar pulau jawa maupun dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 68,56, TOEFL sebesar 412,08, IPK sebesar 3,36 dengan lama studi kurang dari 4 tahun yaitu 3 tahun kurang 11 bulan.

- Cluster* 3 memiliki jumlah data sebanyak 27 data dan mayoritas berasal dari SMA, Asal 2 tetapi luar pulau jawa serta memiliki rata-rata nilai matematika sebesar 69,72, TOEFL sebesar 419,11, IPK sebesar 3,46 dengan lama studi 3 tahun 9 bulan.

Hasil dari evaluasi pola pada dataset Program Studi T dengan jumlah *cluster* sebanyak 4 *cluster*, maka diperoleh *cluster* 0 sebagai *cluster* yang terbaik untuk rekomendasi tepat waktu jika nilai rata-rata matematika besar sama dengan 73,89 dengan asal dari dalam pulau jawa dan rata-rata IPK sebesar 3,62.

4. Pada dataset Program Studi U dengan jumlah 90 data diperoleh hasil pengujian yang disajikan pada Tabel 12.

Tabel 12. Hasil Pengujian Dataset Program Studi U

Nilai K (Jumlah Cluster)	Hasil Silhouette Coefficient
2	0,65
3	0,44
4	0,23

Berdasarkan hasil pengujian pada dataset Program Studi U dapat ditarik kesimpulan bahwa jumlah *cluster* terbaik pada dataset Program Studi U yaitu 2. Sehingga diperoleh 2 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut:

- Cluster* 0 memiliki jumlah data sebanyak 64 data dan mayoritas berasal dari SMA, Asal 2 dalam pulau jawa serta memiliki rata-rata nilai matematika sebesar 82,24, TOEFL sebesar 412,44, IPK sebesar 3,49 dengan lama studi 3 tahun 11 bulan.
- Cluster* 1 memiliki jumlah data sebanyak 26 data dan mayoritas berasal dari SMA Asal 2 dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 82,05, TOEFL sebesar 434,15, IPK sebesar 3,25 dengan lama studi lebih dari 4 tahun yaitu 4 tahun 8 bulan.

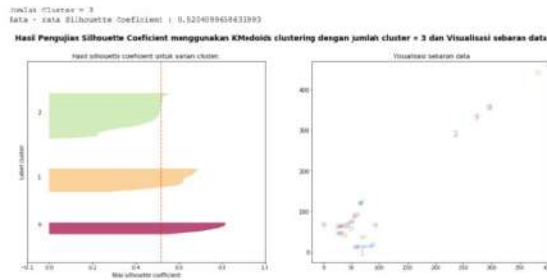
Hasil dari evaluasi pola pada dataset Program Studi U dengan 2 jumlah *cluster*, maka diperoleh *cluster* 0 sebagai *cluster* yang terbaik untuk rekomendasi tepat waktu jika nilai rata-rata matematika besar sama dengan 82,24 dengan asal dari dalam pulau jawa dan rata-rata IPK sebesar 3,49.

DISKUSI

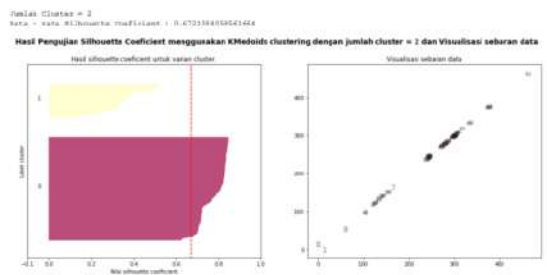
Penelitian dari dataset setiap prodi menghasilkan cluster-cluster yang direkomendasikan berdasarkan hasil pengujian data menggunakan metode *Silhouette Coefficient* dan visualisasi sebaran data yang telah diproses. Hasil dari pengujian tersebut memiliki range nilai dari -1 sampai 1. Jika hasil pengujian bernilai -1 atau mendekati nilai tersebut, maka dapat dikatakan bahwa nilai tersebut tidak bagus dan sebaliknya. Jika hasil pengujian bernilai 1 atau mendekati nilai maka dapat disimpulkan bahwa hasil pengujannya tersebut bagus [12].

Gambar 3, 4, 5 dan 6 merupakan hasil pengujian yang dilakukan pada setiap dataset program studi, nilai diperoleh dari hasil pengujian diatas rata-rata yang diikuti dengan tingkat ketebalan plot siluet yang cukup besar. Sehingga dapat disimpulkan hasil pengujian yang dilakukan pada setiap program studi sudah bagus, karna dapat dilihat dari hasil pengujian bahwa semua hasil pengujian memiliki nilai yang mendekati 1 dan hanya pada data program studi T yang sedikit memiliki nilai negatif. Tingkat keberagaman data dan

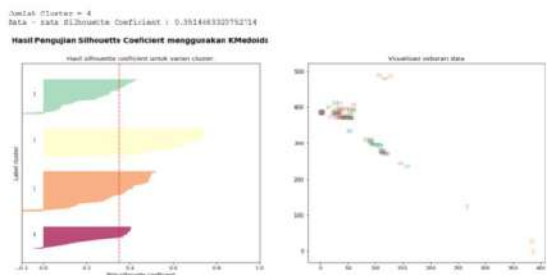
kemiripan data khususnya pada nilai atribut sangat mempengaruhi kedekatan data tersebut dalam sebuah *cluster*.



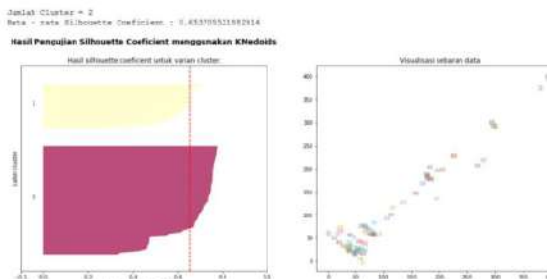
Gambar 3. Hasil Pengujian & Visualisasi Sebaran Data Program Studi R



Gambar 4. Hasil Pengujian & Visualisasi Sebaran Data Program Studi S



Gambar 5. Hasil Pengujian & Visualisasi Sebaran Data Program Studi



Gambar 6. Hasil Pengujian & Visualisasi Sebaran Data Program Studi

KESIMPULAN DAN SARAN

Berdasarkan hasil dari penelitian ini dapat ditarik kesimpulan, sebagai berikut:

- Penelitian menghasilkan beberapa data mahasiswa yang direkomendasikan dan semuanya berasal dari Pulau Jawa. Pada

Herri Kumia

dataset program studi R mahasiswa yang di rekomendasikan berjumlah 9 dengan *cluster* yang terbentuk sebanyak 3 *cluster*, dataset program studi S mahasiswa yang di rekomendasikan berjumlah 57 dengan *cluster* yang terbentuk sebanyak 2 *cluster*, dan dataset program studi U mahasiswa yang di rekomendasikan berjumlah 64 dengan *cluster* yang terbentuk sebanyak 2 *cluster*. Program studi R, S, dan U yang memiliki rata-rata nilai matematika minimal 82. Sedangkan untuk dataset program studi T diperoleh 35 data mahasiswa yang direkomendasikan dengan *cluster* yang terbentuk 4 *cluster* dan memiliki rata-rata nilai matematika sebesar 73,89.

- Hasil pengujian dengan menggunakan metode *Silhouette Coefficient* memperoleh hasil yang cukup bagus dengan nilai *silhouette coefficient* untuk dataset Program Studi R sebesar 0.52, untuk dataset Program Studi S sebesar 0.67, untuk dataset Program Studi T sebesar 0.35 dan untuk dataset Program Studi U sebesar 0.65.

Pada penelitian ini masih memiliki kekurangan, maka perlu adanya pengembangan untuk kedepannya untuk memperoleh hasil yang lebih bagus. Saran dari penulis yaitu perlu melakukan penambahan dataset dan atribut data, sehingga dapat meningkatkan akurasi dari pengelompokan dalam penelitian selanjutnya.

REFERENSI

- W. Widarto, "Faktor Penghambat Studi Mahasiswa yang Tidak Lulus Tepat Waktu di Jurusan Pendidikan Teknik Mesin FT UNY," *J. Din. Vokasional Tek. Mesin*, vol. 2, no. 2, p. 127, 2017.
- D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokan Wilayah Sebaran Cacat pada Anak," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 64, 2018.
- D. F. Pramesti, M. Tanzil Furqon, and C. Dewi, "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 723–732, 2017.
- L. Zahrotun, "Analisis Pengelompokan Jumlah Penumpang Bus Trans Jogja Menggunakan metode Clustering K-Means dan Agglomerative Hierarchical Clustering (AHC)," *J. Inform.*, vol. 9, no. 1, pp. 1039–1047, 2015.
- L. Zahrotun, N. hutami Putri, and A. N. Khusna, "The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesisi Titles," in *12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2018.
- L. Zahrotun, "Text Mining for Internship Titles Clustering Using Shared Nearest-Neighbor Method," *Comput. Eng. Appl.*, vol. 6, no. 3, 2017.
- S. Defiyanti, M. Jajuli, and N. Rohmawati, "Optimalisasi K-MEDOID dalam Pengklasteran Mahasiswa Pelamar Beasiswa dengan CUBIC CLUSTERING CRITERION," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 1, pp. 211–218, 2017.
- N. K. Kaur, U. Kaur, and D. Singh, "K-Medoid Clustering Algorithm- A Review," *Int. J. Comput. Appl. Technol.*, vol. 1, no. 1, pp. 42–45, 2014.
- R. Handoyo, R. Rumani, and S. M. Nasution, "Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K-Means Pada Pengelompokan

- Dokumen," *JSM STMIK Mikroskil*, vol. 15, no. 2, pp. 73–82, 2014.
- [10] M. B. Al- Zoubi and M. A. Rawi, "An Efficient Approach for Computing Silhouette Coefficients," *J. Comput. Sci.*, vol. 4, no. 3, pp. 252–255, 2008.
- [11] J. Han, Jiawei; Kamber, Micheline; Pei, *Data mining: Data mining concepts and techniques*, 2014.
- [12] W. Bagus, A. E. Budianto, and A. S. Wiguna, "Implementasi Metode K-Medoids Clustering untuk Mengetahui Pola Pemilihan Program Studi," *J. Terap. Sains Teknol.*, vol. 1, no. 3, pp. 54–69, 2019.
- [13] G. S. Nugraha, Hairani, and R. F. P. Ardi, "Aplikasi Pemetaan Kualitas Pendidikan Di Indonesia Menggunakan Metode K-Means," *J. MATRIK*, vol. 17, no. 2, pp. 13–23, 2018.

BIOGRAFI PENULIS



Herri Kurnia

Salah satu mahasiswa Universitas Ahmad Dahlan (UAD). Jurusan / Program Studi Teknik Informatika. Diterima menjadi mahasiswa Program Studi Teknik Informatika pada tahun 2016 di UAD.



Lisna Zahrotun, S.T., M.Cs.

Bekerja sebagai dosen Program Studi Teknik Informatika Fakultas Teknologi Industri di Universitas Ahmad Dahlan sampai sekarang. Bidang keahlian data mining, text mining, sistem pendukung keputusan.



Utaminingsih Linarti, S.T., M.T.

Bekerja sebagai dosen Program Studi Teknik Industri Fakultas Teknologi Industri di Universitas Ahmad Dahlan sampai sekarang. Bidang keahlian data rantai pasok, learning system.

TATA NAMA

Arti atau maksud dari symbol yang ada pada persamaan-persamaan yang digunakan.

- Σ berarti jumlah atau total keseluruhan
- $| |$ berarti Absolute atau mutlak
- ϵ berarti Elemen atau anggota
- n berarti jumlah data atau banyak data

HASIL CEK_Kurnia, Zahrotun, Linarti_K-Medoids, Silhouette Coefficient, One Hot Encoding, Euclidean Distance, Cluster

ORIGINALITY REPORT

3%

SIMILARITY INDEX

3%

INTERNET SOURCES

2%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1

www.scribd.com

Internet Source

3%

Exclude quotes On

Exclude matches < 2%

Exclude bibliography On



InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan

ISSN (Print) 2540-7597 | ISSN (Online) 2540-7600



Available online at : <http://bit.ly/InfoTekJar>

Pengelompokan Mahasiswa Berdasarkan Data Akademik Sebelum Kuliah dan Masa Studi Menggunakan *K-Medoids*

Herri Kurnia, Lisna Zahrotun, Utaminingsih Linarti

Universitas Ahmad Dahlan Fakultas Teknologi Industri Program Studi Teknik Informatika, Jl. Ringroad Selatan, Yogyakarta, 55191., Indonesia

KEYWORDS

K-Medoids, Silhouette Coefficient, One Hot Encoding, Euclidean Distance, Cluster

CORRESPONDENCE

Phone: +62 81328265007

E-mail: lisna.zahrotun@tif.uad.ac.id

ABSTRACT

The research objective is to obtain information about the results of grouping that are useful for the campus, especially study programs, to be used as consideration for future admissions by grouping, causing a mismatch between the number of students and the existing campus facilities. This research was conducted at university X which has several faculties, one of which is faculty Y which consists of the R study program, S study program, T study program and U study program. This research uses the K-Medoids method. The stages of this research started with load dataset, data cleaning, data selection, data transformation with one hot encoding, euclidean distance, and k-medoids to produce clusters. Testing the quality of the clusters in this study using the silhouette coefficient. The research resulted in recommended student data and all of them came from Java Island. In the dataset of study programs R, S, and U, the recommended data are obtained with a total number of 9, 57, and 64, respectively, which have an average math score of at least 82. Meanwhile, for the T study program dataset, there are 35 data with an average mathematical value. amounted to 73.89. The test results for the dataset of study programs R, S, T and U are 0.52, 0.67, 0.35, and 0.65 respectively, so the results are quite good.

ABSTRAK

Tujuan penelitian yaitu untuk memperoleh informasi tentang hasil pengelompokan yang berguna bagi pihak kampus terutama prodi untuk dijadikan bahan pertimbangan untuk penerimaan mahasiswa baru kedepannya dengan cara mengelompokkan, sehingga menyebabkan ketidaksesuaian antara jumlah mahasiswa dengan fasilitas kampus yang ada. Penelitian ini dilakukan pada universitas X memiliki beberapa fakultas salah satunya fakultas Y yang terdiri dari program studi R, program studi S, program studi T dan program studi U. Penelitian ini menggunakan metode *K-Medoids*. Tahapan penelitian ini dimulai dari *load* dataset, data *cleaning*, data *selection*, transformasi data dengan *one hot encoding*, *euclidean distance*, dan *k-medoids* untuk menghasilkan *cluster*. Pengujian kualitas *cluster* dalam penelitian ini menggunakan *silhouette coefficient*. Penelitian menghasilkan data mahasiswa yang direkomendasikan dan semuanya berasal dari Pulau Jawa. Pada dataset program studi R, S, dan U diperoleh data yang direkomendasikan dengan jumlah berturut-turut 9, 57, dan 64 yang memiliki rata-rata nilai matematika minimal 82. Sedangkan untuk dataset program studi T terdapat 35 data dengan rata-rata nilai matematika sebesar 73,89. Hasil pengujian untuk dataset program studi R, S, T dan U berurut-turut sebesar 0,52, 0,67, 0,35, dan 0,65 sehingga dinyatakan hasilnya cukup bagus.

PENDAHULUAN

Universitas X mempunyai beberapa fakultas, salah satunya Fakultas Y. Fakultas Y memiliki 4 program studi yaitu program studi R, program studi S, program studi T dan program studi U. Peningkatan jumlah mahasiswa baru terjadi setiap tahun untuk semua program studi di Fakultas Y. Namun peningkatan tersebut tidak diimbangi dengan prosentase jumlah kelulusan mahasiswa tepat waktu di setiap program studi, yang secara agregat menjadi prosentase kelulusan mahasiswa di Fakultas Y. Salah satu faktor yang menyebabkan rendahnya prosentase kelulusan mahasiswa tepat waktu antara lain : (1) kurang minat pada pilihan prodi, (2) masih mengikuti kuliah, (3) metode bimbingan yang tidak intensif, (4) kurang lengkapnya fasilitas kampus dan (5) pengaruh lingkungan[1].

Tabel 1. Data Jumlah Mahasiswa dan Kelulusan Fakultas Y

Tahun	Jumlah Mahasiswa	Jumlah Lulusan
2012	360	14 atau 4%
2013	340	22 atau 6%
2014	591	55 atau 9%
2015	837	181 atau 22%

Adanya ketidakseimbangan tersebut akan berimplikasi pada faktor lain dalam pembelajaran akademik, misalnya (1) tidak seimbang jumlah dosen dan mahasiswa, (2) meningkatnya penggunaan fasilitas kampus, seperti laboratorium yang tidak sesuai dengan jumlah mahasiswa sehingga membutuhkan penambahan fasilitas atau penambahan jam slot praktikum. Selama ini data mahasiswa sebelum masuk seperti nilai raport, asal sekolah dan asal kabupaten dengan data kelulusan mahasiswa yaitu lama studi, TOEFL dan IPK belum pernah dikelompokkan berdasarkan aturan tertentu, hal ini dijadikan bahan pertimbangan bagi pihak Fakultas Y khususnya setiap program studi dalam menyeleksi mahasiswa baru.

Data yang digunakan merupakan data yang cukup besar karena mencakup semua mahasiswa angkatan 2014 – 2015 dari program studi R, S, T dan U. *Data Mining* merupakan suatu teknik yang diperuntukan untuk memproses penggalan informasi yang belum diketahui dalam suatu dataset atau basis data yang besar [2]. *Data Mining* mempunyai banyak teknik pengolahan dalam mengolah data, salah satu tekniknya yaitu Teknik *Clustering*.

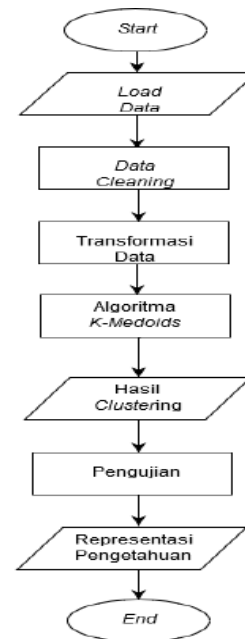
Clustering adalah sebuah teknik untuk memproses data dan mengelompokkan data dalam suatu kelas yang memiliki kemiripan objek [3]. Pada Teknik *Clustering* terdapat beberapa algoritma yang dapat mengelompokkan data dengan baik, diantaranya K-Means [4] [5], Shared Nearest Neighbour (SNN) [6] dan Algoritma *K-Medoids*. Penelitian ini menggunakan Metode *K-Medoids* dengan pendekatan *Euclidean Distance*. Algoritma *K-Medoids* efektif untuk mengatasi dataset yang mana ada beberapa data (outlier) yang terletak sangat jauh dari kebanyakan data lainnya [7]. Pengolahan data pada penelitian ini menggunakan Algoritma *K-Medoids* dengan pendekatan *Euclidean Distance* dengan data yang digunakan bertipe numerik dan teks.

Berdasarkan permasalahan yang ada, disimpulkan bahwa untuk mengatasi masalah mahasiswa yang belum lulus tepat waktu maka diperlukan pengelompokkan menggunakan Algoritma *K-Medoids* berdasarkan ketentuan – ketentuan tertentu. Sehingga hasil

pengelompokkan tersebut dapat memberikan informasi dan juga menjadi bahan pertimbangan dalam penyeleksi calon mahasiswa baru yang diharapkan membantu Fakultas Y khususnya setiap program studi yang ada di fakultas tersebut.

METODE

Metode yang digunakan dalam penelitian ini adalah metode *K-Medoids*. *K-Medoids* adalah algoritma *clustering* yang menggunakan objek sebagai pusat *cluster* pada setiap *cluster*. Algoritma *K-Medoids* menggunakan objek sebagai perwakilan (medoid) sebagai pusat cluster untuk setiap cluster, sedangkan K-Means menggunakan nilai rata-rata (mean) sebagai pusat cluster [8]. Tahapan pada penelitian ini melalui beberapa tahap yang dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

Penjelasan dari tahap-tahap dalam penelitian ini, sebagai berikut

1. Load Data

Load Data merupakan proses mengambil data dari dataset yang sudah ditentukan berdasarkan kebutuhan sedemikian rupa

2. Data Cleaning

Data Cleaning adalah sebuah proses pembersihan data dengan cara memilih data atau menghapus data yang sesuai dengan ketentuan yang bertujuan untuk menghindari data yang *noise*.

3. Transformasi Data

Transformasi Data merupakan teknik untuk mengubah data menjadi bentuk lain yang diasumsikan memenuhi analisis ragam. Teknik transformasi yang digunakan pada penelitian ini menggunakan Teknik *One Hot Encoding*. *One Hot Encoding* adalah sebuah cara untuk mentransformasikan atau merepresentasikan data agar dapat dipahami komputer. Prinsip kerja metode ini yaitu dengan membuat sebuah array 1 dimensi yang memiliki panjang sebanyak jenis class yang ada dan memiliki nilai biner [9].

4. Algoritma *K-Medoids*

Data akan diambil secara acak untuk dijadikan data pusat pada *cluster*, setiap data berpeluang menjadi data pusat tetapi data yang paling tengahnya yang dijadikan data pusat pada suatu *cluster*

berdasarkan ketentuan dari Algoritma *K-Medoids*. Langkah – langkah Algoritma *K-Medoids*, sebagai berikut :

- a. Inisialisasi pusat cluster sebanyak k (jumlah cluster).
- b. Kelompokkan setiap data ke cluster terdekat menggunakan pendekatan *Euclidian Distance* untuk menghitung jarak antar data dengan persamaan:

$$d(x, y) = ||x - y|| = \sqrt{(\sum_{i=1}^n (x_{(i)} - y_{(i)}))^2}; 1, 2, 3, \dots, n \dots (1)$$

Keterangan :

$x_{(i)}$ = data ke i pertama.

$y_{(i)}$ = data ke i kedua.

n = banyak data.

c. Kemudian pilih data secara random pada setiap cluster yang dijadikan calon medoid baru.

d. Setelah itu, hitung jarak setiap data yang berada pada masing-masing cluster dengan calon medoid baru.

e. Kemudian menghitung total simpangan (S) dengan menghitung nilai total distance baru – total distance lama. Jika $S < 0$, maka ganti objek dengan data *cluster* untuk membentuk sekumpulan k objek baru sebagai *medoid*.

f. Ulangi langkah 3 sampai 5 sampai tidak terjadi perubahan *medoid*, sehingga didapatkan cluster beserta anggota cluster masing-masing.

5. Hasil *Clustering*

Hasil *Clustering* merupakan hasil pengolahan data dari tahap sebelumnya yaitu tahap Algoritma *K-Medoids*, dimana hasil ini berupa data-data yang sudah dikelompokkan berdasarkan tingkat kemiripan antar data.

6. Pengujian

Pada penelitian ini, pengujian menggunakan Metode *Silhouette Coefficient*, yang mana metode ini akan menghitung tingkat kedekatan yang terdapat antar data atau objek dalam suatu *cluster*. Langkah – langkah pada proses *silhouette coefficient* [9], sebagai berikut :

- a. Hitung rata-rata jarak dari suatu dokumen misalkan i dengan semua dokumen lain yang berada dalam satu *cluster* [10].

$$a(i) = \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i, j) \dots\dots\dots (2)$$

Yang mana j merupakan dokumen lain dalam cluster A dan d(i,j) bermakna jarak antar dokumen i dengan j.

- b. Hitung rata-rata jarak dari dokumen i tersebut dengan semua dokumen di cluster lain, dan diambil nilai terkecilnya [10].

$$d(i, C) = \frac{1}{|A|} \sum_{j \in C} d(i, j) \dots\dots\dots (3)$$

dengan d(i,C) adalah jarak rata-rata dokumen i dengan semua objek pada *cluster* lain C dimana $A \neq C$ [10].

$$b(i) = \min_{C \neq A} d(i, C) \dots\dots\dots (4)$$

- c. Nilai *Silhouette Coefficient* [11].

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \dots\dots\dots (5)$$

Keterangan :

$s(i)$ = Nilai *Silhouette Coefficient*.

$a(i)$ = Rata-rata jarak i terhadap semua objek di klaster A.

$b(i)$ = Rata-rata jarak i terhadap semua objek klaster lain.

7. Representasi Pengetahuan

Tahap terakhir ini merupakan tahap pengambilan kesimpulan berdasarkan dari hasil pengujian. Hasil pengujian akan berada pada kisaran -1 sampai 1. Jika nilai dari hasil pengujian mendekati nilai 1, maka suatu cluster yang telah dilakukan pengelompokkan data dapat dikatakan baik. Sedangkan jika nilai dari hasil pengujian mendekati nilai -1, maka dapat disimpulkan bahwa *cluster* tersebut buruk dalam mengelompokkan data [12].

HASIL

Penelitian ini menggunakan data mahasiswa dari program studi R, S, U dan T dengan jumlah dari masing-masing data berturut-turut yaitu 31, 90, 90, dan 76. Data tersebut merupakan data mahasiswa angkatan 2014 – 2015 dengan beberapa atribut yaitu Angkatan, NIM, Nama, Prodi, Jalur Masuk, Sekolah, Nama Sekolah, Asal Kabupaten, Wilayah I, Wilayah II, Wilayah III, MTK, LAMA.

Tabel 2. Load Data

Angkatan	NIM	Nama	Prodi	Jalur_Masuk	Sekolah	Nama_Sekolah	Asal_Kabupaten	Wilayah I	Wilayah II	Wilayah III	MTK	LAMA
0	2014 1400019002	MUHAMMAD ANDREANI PRATAMA	TEKNIK INDUSTRI	PMDK-Rapor1	SMK	SMK Negeri 2 Yogyakarta	Di Yogyakarta	0	1	0	42	550000
1	2014 1400019005	DEDI MUSTAL	TEKNIK INDUSTRI	PMDK-Rapor1	SMK	SMK KHARAYA DARMA 1 KOTABUMI	Lampung	0	1	0	55	000000
2	2014 1400019012	NOVRIMAWAN	TEKNIK INDUSTRI	PMDK-Rapor1	SMA	SMA Negeri 12 Weraing	Jambi	0	1	0	87	566667
3	2014 1400019014	SANDIJI SAIWIKO PRHATIMOKO	TEKNIK INDUSTRI	PMDK-Rapor1	SMK	SMK Negeri 3 Yogyakarta	Di Yogyakarta	0	1	0	80	333333
4	2014 1400019017	MUHAMMAD KRISHNA PUTRA	TEKNIK INDUSTRI	PMDK-Rapor1	SMA	smp negeri 2 bandar lampung	Lampung	0	1	0	95	333333
71	2015 1500019163	WAHDI LUTHFI RAMADHANI	TEKNIK INDUSTRI	PMDK-Rapor1	SMA	SMA NEGERI 5 TEDO	Jambi	0	1	0	85	333333
72	2015 1500019165	BITAN POFANI	TEKNIK INDUSTRI	PMDK-Rapor1	SMA	SMA Negeri 1, Sidorong	Jawa Tengah	0	1	0	84	566667
73	2015 1500019166	RAMA YUDHI FERIMENDO	TEKNIK INDUSTRI	PMDK-Rapor1	SMA	SMA Budi Utomo, Parak	Jawa Timur	0	1	0	84	000000
74	2015 1500019205	SANIA LUNA WAKHU ELLENORA	TEKNIK INDUSTRI	PMDK-Rapor1	SMA	SMA Negeri 1 Tembungan Hulu	Riau	0	1	0	85	000000
75	2015 1500019207	DEA AGRIYAH AMELIA	TEKNIK INDUSTRI	PMDK-Rapor1	SMA	SMA NEGERI 2 CIREBON	Jawa Barat	0	1	0	87	000000

76 rows x 13 columns

Tahap selanjutnya yaitu melakukan proses data *cleaning* tetapi untuk atribut NIM dan Nama Sekolah disimpan. selanjutnya transformasi data menggunakan metode *One Hot Encoding* pada atribut Sekolah. Pada atribut asal kabupaten, data ditransformasi menjadi 3 yaitu asal 1 yang berwarna abu-abu terdiri Maluku Utara dan Kalimantan Tengah, asal 2 yang berwarna coklat terdiri dari Pulau Jawa, Pulau Sumatra, Pulau Sulawesi, Kalimantan Barat, Kalimantan Selatan, Bali, NTT dan NTB, asal 3 yang berwarna hijau terdiri dari Pulau Papua, Kalimantan Timur dan Maluku. Pengelompokkan ini berdasarkan tingkat kualitas pendidikan dengan parameter tertentu.



Gambar 2. Pemetaan Kualitas Pendidikan di Indonesia[13]

Atribut asal sekolah diubah menjadi 3 atribut baru yaitu wilayah I , wilayah II, dan wilayah III yang disisipkan kedalam dataset beserta nilai yang sudah diperoleh dari hasil transformasi berdasarkan pemetaan kualitas pendidikan di Indonesia.

Tabel 3. Hasil Pemrosesan Data *Cleaning & Transformasi Data*

Wilayah I	Wilayah II	Wilayah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	
0	0	1	0	42.550000	1748	3.42	466	0	0	1
1	0	1	0	56.000000	1749	2.91	470	0	0	1
2	0	1	0	87.666667	1513	3.09	413	0	1	0
3	0	1	0	80.333333	1513	3.34	410	0	0	1
4	0	1	0	86.333333	1842	3.22	456	0	1	0
...
71	0	1	0	88.333333	1455	3.48	406	0	1	0
72	0	1	0	84.666667	1455	3.66	413	0	1	0
73	0	1	0	84.000000	1478	3.30	463	0	1	0
74	0	1	0	85.000000	1455	3.44	406	0	1	0
75	0	1	0	87.000000	1455	3.57	436	0	1	0

76 rows x 10 columns

Setelah didapatkan hasil pemrosesan tersebut, selanjutnya masuk ketahap Algoritma *K-Medoids*, dimana setiap data dihitung jarak kedekatannya dengan data lain dengan menggunakan metode *Euclidean Distance*.

Data ke -1,2 :

$$d(1,2) = \sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (42.55-56.0)^2 + (1748-1749)^2 + (3.42-2.91)^2 + (466-470)^2 + (0-0)^2 + (0-0)^2 + (1-1)^2} = 14.077$$

⋮

Data ke - 1, 76 :

$$d(1,76) = \sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (42.55-87.0)^2 + (1748-1455)^2 + (3.42-3.57)^2 + (466-436)^2 + (0-0)^2 + (0-1)^2 + (1-0)^2} = 297.87$$

Hasil dari perhitungan jarak setiap data / *Euclidean Distance* tersebut yang akan menjadi dataset awal untuk melakukan proses *K-Medoids*.

Tabel 4. Hasil Perhitungan *Euclidean Distance*

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	...	D67	D68
D1	0.00000	14.077024	245.095100	244.517048	104.187429	239.702741	240.420002	164.696915	60.238612	142.160679	...	295.365748	295.459971
D2	14.077024	0.000000	244.946910	244.720693	98.829162	239.870459	240.460121	164.670766	54.991899	141.626794	...	295.368904	294.937071
D3	245.095100	244.946910	0.000000	6.052046	331.800836	37.006234	6.708390	90.227667	278.617145	116.803254	...	42.094199	83.802889
D4	244.517048	244.720693	6.052046	0.000000	332.257462	40.576580	6.310133	88.961380	278.643338	116.486690	...	42.510172	85.642638
D5	104.187429	98.829162	331.800836	332.257462	0.000000	330.055382	331.466664	246.516062	56.200648	220.298205	...	373.483793	387.428017
...
D72	302.567542	302.821485	58.425942	58.702807	580.224818	72.815772	58.220099	147.329657	336.911713	174.172022	...	17.619815	67.518200
D73	300.722249	300.847703	58.082531	58.257618	309.385376	68.857766	58.157290	147.636250	336.173984	174.484326	...	16.057565	60.181749
D74	273.183360	272.636515	61.142161	60.638258	364.074767	37.456727	58.646986	138.103797	312.924917	160.621492	...	50.527219	25.387074
D75	302.067120	302.283445	58.482763	58.341990	330.216946	72.828888	58.950836	147.183714	336.616003	174.114525	...	17.482901	67.183529
D76	297.870484	297.882663	62.388318	63.926717	387.517183	59.666762	61.684327	161.464188	338.173292	177.722825	...	28.093207	37.656723

76 rows x 76 columns

Data hasil perhitungan *Euclidean Distance* diproses dengan algoritma *K-Medoids* dan menghasilkan beberapa *cluster* beserta anggota – anggota yang mirip dengan *medoid* atau data pusatnya yang menjadi hasil akhir dari clustering. Data yang disimpan diawal tadi yaitu data NIM dan Nama Sekolah dikembalikan sesuai masing – masing data.

Tabel 5. Hasil *Clustering* Program Studi S

NIM	Nama_Sekolah	Wilayah I	Wilayah II	Wilayah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	Cluster	
0	1400019002	SMK Negeri 2, Yogyakarta	0	1	0	42.550000	1748	3.42	466	0	0	1	1
1	1400019008	SUK KHARVA DHARMA 1 KOTABUNING	0	1	0	56.000000	1749	2.91	470	0	0	1	1
2	1400019012	SMA Negeri 12 Merangin	0	1	0	87.666667	1513	3.09	413	0	1	0	0
3	1400019014	SMK Negeri 3, Yogyakarta	0	1	0	80.333333	1513	3.34	410	0	0	1	0
4	1400019017	sma pentil 2 bandar lampung	0	1	0	86.333333	1842	3.22	456	0	1	0	1
...
71	15000019163	SMA NEGERI 5 TEBU	0	1	0	88.333333	1455	3.48	406	0	1	0	0
72	15000019165	SMA Negeri 1, Bandungan	0	1	0	84.666667	1455	3.66	413	0	1	0	0
73	15000019166	SMA Budi Utomo, Perak	0	1	0	84.000000	1478	3.30	463	0	1	0	0
74	1500019295	SMA Negeri 1 Tembloran Huru	0	1	0	85.000000	1455	3.44	406	0	1	0	0
75	1500019297	SMA NEGERI 2 CREBON	0	1	0	87.000000	1455	3.57	436	0	1	0	0

76 rows x 13 columns

Setiap Dataset melalui proses yang sama seperti paparan diatas dan menghasilkan hasil akhir yang berupa data-data yang sudah di-*clustering*.

Tabel 6. Hasil *Clustering* Program Studi R

NIM	Nama_Sekolah	Wilayah I	Wilayah II	Wilayah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	Cluster
0 1400022003	SMK Negeri 1 Batam	0	1	0	52.87	1455	3.41	406	0	0	1	2
1 1400022005	SMK Negeri 2, Pekanbaru	0	1	0	81.00	1392	3.32	406	0	0	1	1
2 1400022010	SMA Negeri 2, Kalilanda	0	1	0	79.87	1679	3.19	470	0	1	0	0
3 1400022024	SMK Hassanah Pekanbaru	0	1	0	78.87	1728	3.34	416	0	0	1	0
4 1400022032	sma negeri 1 bandar sribhawono	0	1	0	78.87	1513	3.44	420	0	1	0	2
5 1400022039	SMA NEGERI 1 CIAWIGEBANG	0	1	0	78.87	1451	3.79	440	0	1	0	2
6 1400022042	SMA Negeri 1, Marabahan	0	1	0	84.87	1455	3.86	426	0	1	0	2
7 1400022046	SMA N 2 FAKFAK	0	0	1	88.33	1748	3.02	426	0	1	0	0
8 1400022048	SMA Negeri 1, Cibarusah	0	1	0	77.33	1513	3.21	426	0	1	0	2
9 1400022056	smk muh 1 bambanglipuro	0	1	0	81.33	1513	3.35	426	0	0	1	2
10 1400022058	SMK Negeri 2, Yogyakarta	0	1	0	83.33	1392	3.74	420	0	0	1	1
11 1400022079	SMA N 1 TEMPEL	0	1	0	80.33	1834	3.31	443	0	1	0	0
12 1400022080	smk muhammadiyah 2 mentoyudan	0	1	0	80.87	1749	3.20	416	0	0	1	0
13 1500022003	SMA Negeri 2, Ketapang	0	1	0	78.87	1398	3.85	473	0	1	0	1
14 1500022008	SMA Negeri 1, Keruak	0	1	0	78.87	1455	3.62	450	0	1	0	2
15 1500022010	SMA Negeri 1, Temate	1	0	0	80.00	1471	3.26	456	0	1	0	2
16 1500022019	SMA Muhammadiyah, Wonosobo	0	1	0	82.87	1404	3.32	410	0	1	0	1
17 1500022027	SMA Negeri 3, Slawi	0	1	0	79.87	1439	3.62	406	0	1	0	2
18 1500022031	MA Negeri, Ketapang	0	1	0	77.87	1455	3.63	413	1	0	0	2
19 1500022033	SMK Negeri 1 Sambang	0	1	0	88.33	1378	3.37	420	0	0	1	1
20 1500022034	SMK N 1 BLORA	0	1	0	84.87	1378	3.68	400	0	0	1	1
21 1500022035	SMA Negeri 2, Balikpapan	0	0	1	77.87	1432	3.82	420	0	1	0	2
22 1500022037	SMA Negeri 1, Sengah Temila	0	1	0	79.33	1455	3.87	446	0	1	0	2
23 1500022042	SMA NEGERI 12 MERANGIN	0	1	0	84.87	1471	3.48	446	0	1	0	2
24 1500022047	SMK Negeri 1, Sedayu	0	1	0	81.87	1455	3.36	403	0	0	1	2
25 1500022053	SMA Negeri 2, Dumai	0	1	0	83.33	1455	3.64	416	0	1	0	2
26 1500022058	SMA Negeri 1, Kebumen	0	1	0	80.00	1404	3.66	403	0	1	0	1
27 1500022059	SMA Negeri 1, Manna	0	1	0	85.00	1439	3.51	440	0	1	0	2
28 1500022067	SMK NEGERI 1 TONJONG	0	1	0	81.33	1404	3.59	443	0	0	1	1
29 1500022069	SMAAN 4 MUARO JAMBI	0	1	0	88.33	1404	3.41	413	0	1	0	1
30 1500022083	SMK Negeri 2 Purwokerto	0	1	0	75.87	1439	3.65	406	0	0	1	2

Tabel 7. Hasil *Clustering* Program Studi T

NIM	Nama_Sekolah	Wilayah I	Wilayah II	Wilayah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	Cluster
0 1400020002	SMA Negeri 5, Manna	0	1	0	54.666667	1368	3.80	456	0	1	0	0
1 1400020011	SMA N 1 MLONGGO	0	1	0	84.666667	1749	3.52	400	0	1	0	1
2 1400020016	STN Perindustrian, Yogyakarta	0	1	0	58.333333	1368	3.80	456	0	0	1	0
3 1400020017	MA Negeri, Praya	0	1	0	58.333333	1368	3.80	456	1	0	0	0
4 1400020018	SMA N 1 WITIA PONDIA	0	1	0	55.666667	1368	3.80	456	0	1	0	0
...
85 1500020149	SMA Muhammadiyah 1, Yogyakarta	0	1	0	81.666667	1377	3.51	423	0	1	0	3
86 1500020154	MA Negeri 2, Banjarnegara	0	1	0	82.000000	1377	3.67	423	1	0	0	3
87 1500020185	SMA Negeri 1, Dayeuhluhur	0	1	0	58.000000	1355	3.84	463	0	1	0	0
88 1500020179	SMA Alabio, Alabio	0	1	0	91.333333	1355	3.74	423	0	1	0	3
89 1500020183	SMA Negeri 2, Muara Bungo	0	1	0	78.500000	1455	3.39	470	0	1	0	1

90 rows x 13 columns

Tabel 8. Hasil *Clustering* Program Studi U

NIM	Nama_Sekolah	Wilayah I	Wilayah II	Wilayah III	MTK	LAMA_STUDI	IPK	TOEFL	MA	SMA	SMK	Cluster
0 1400018012	SMA Gedjah Mada, Yogyakarta	0	1	0	83.333333	1455	3.46	400	0	1	0	0
1 1400018016	MA Negeri 3, Yogyakarta	0	1	0	77.666667	1455	3.26	400	1	0	0	0
2 1400018017	SMA Negeri 1, Babakan	0	1	0	89.333333	1513	3.57	408	0	1	0	0
3 1400018026	SMK TARUNA BANGSA CIMIS	0	1	0	78.333333	1830	2.66	403	0	0	1	1
4 1400018041	SMAAN 01 WITIA PONDIA	0	1	0	79.000000	1613	3.63	400	0	1	0	0
...
85 1500018242	MA Negeri 1, Metro	0	1	0	79.000000	1455	3.75	423	1	0	0	0
86 1500018246	SMA Negeri 1, Sungapenuh	0	1	0	82.666667	1308	3.76	403	0	1	0	0
87 1500018249	SMA Negeri 3, Medan	0	1	0	81.666667	1455	3.62	440	0	1	0	0
88 1500018255	SMA Negeri 1, Purbalingga	0	1	0	76.666667	1343	3.59	433	0	1	0	0
89 1500018310	SMA KH Mustika, Sukamanah	0	1	0	82.000000	1438	2.99	416	0	1	0	0

90 rows x 13 columns

Semua hasil *Clustering* setiap program studi dilakukan analisis evaluasi pola dan representasi pengetahuan sehingga dapat ditarik kesimpulan yang menghasilkan informasi yang bermanfaat untuk pihak program studi. Berdasarkan hasil *clustering* dari setiap dataset, maka dapat dilakukan evaluasi pola sebagai berikut :

1. Pada dataset Program Studi R dengan jumlah 31 data diperoleh hasil pengujian yang disajikan pada Tabel 9.

Tabel 9. Hasil Pengujian Dataset Program Studi R

Nilai K (Jumlah Cluster)	Hasil Silhouette Coefficient
2	0,518
3	0,52
4	0,435

Berdasarkan hasil pengujian pada dataset Program Studi R dapat ditarik kesimpulan bahwa pada dataset Program Studi R jumlah *cluster* terbaik yaitu 3. Sehingga diperoleh 3 *cluster* dengan evaluasi pola setiap *cluster* sebagai berikut :

- a. *Cluster* 0 memiliki jumlah data sebanyak 5 data dan mayoritas berasal dari SMA, Asal 2, luar pulau jawa serta memiliki rata-rata nilai matematika sebesar 81,5 , TOEFL sebesar 434,2, IPK sebesar 3,21 dengan lama studi lebih dari 4 tahun yaitu 4 tahun 9 bulan.
- b. *Cluster* 1 memiliki jumlah data sebanyak 9 data dan berasal dari Asal 2 tetapi dalam pulau jawa dengan mayoritas berasal dari SMK, serta memiliki rata-rata nilai matematika sebesar 82,92 , TOEFL sebesar 420,88 , IPK sebesar 3,547 dengan lama studi kurang dari 4 tahun yaitu 3 tahun 10 bulan.
- c. *Cluster* 2 memiliki jumlah data sebanyak 17 data dan mayoritas berasal dari SMA, Asal 2, luar pulau jawa serta memiliki rata-rata nilai matematika sebesar 78,62 , TOEFL sebesar 426,23 , IPK sebesar 3,541 dengan lama studi 4 tahun 2 hari.

Hasil dari evaluasi pola pada dataset Program Studi R dengan 3 jumlah *cluster*, maka diperoleh *cluster* 1 sebagai *cluster* yang terbaik untuk rekomendasi tepat waktu jika nilai rata-rata matematika besar sama dengan 82,92 dengan asal dari dalam pulau jawa dan rata – rata IPK sebesar 3,547.

2. Pada dataset Program Studi S dengan jumlah 76 data diperoleh hasil pengujian yang disajikan pada Tabel 10.

Tabel 10. Hasil Pengujian Dataset Program Studi S

Nilai K (Jumlah Cluster)	Hasil Silhouette Coefficient
2	0,67
3	0,36
4	0,27

Berdasarkan hasil pengujian pada dataset Program Studi S dapat ditarik kesimpulan bahwa jumlah cluster terbaik pada dataset Program Studi S yaitu 2. Sehingga diperoleh 2 cluster dengan evaluasi pola setiap cluster sebagai berikut:

- a. Cluster 0 memiliki jumlah data sebanyak 57 data dan mayoritas berasal dari SMA, Asal 2 tetapi dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 82,84, TOEFL sebesar 425,16, IPK sebesar 3,47 dengan lama studi 4 tahun lebih 8 hari.
- b. Cluster 1 memiliki jumlah data sebanyak 19 data dan mayoritas berasal dari SMA, Asal 2 tetapi dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 77,15, TOEFL sebesar 438,89, IPK sebesar 3,31 dengan lama studi 4 tahun 8 bulan.

Hasil dari evaluasi pola pada dataset Program Studi S yang memiliki 2 cluster, maka diperoleh cluster 0 sebagai cluster yang terbaik untuk rekomendasi tepat waktu jika nilai rata-rata matematika besar sama dengan 82,84 dengan asal dari dalam pulau jawa dan rata-rata IPK sebesar 3,47.

3. Pada dataset Program Studi T dengan jumlah 90 data diperoleh hasil pengujian yang disajikan pada Tabel 11.

Tabel 11. Hasil Pengujian Dataset Program Studi T

Nilai K (Jumlah Cluster)	Hasil Silhouette Coefficient
2	0,21
3	0,28
4	0,35

Berdasarkan hasil pengujian pada dataset Program Studi T dapat ditarik kesimpulan bahwa jumlah cluster terbaik adalah 4. Sehingga diperoleh 4 cluster dengan evaluasi pola setiap cluster sebagai berikut :

- a. Cluster 0 memiliki jumlah data sebanyak 35 data dan mayoritas berasal dari SMA, Asal 2 tetapi dalam pulau jawa serta memiliki rata-rata nilai matematika sebesar 73,89, TOEFL sebesar 456,51, IPK sebesar 3,62 dengan lama studi 3 tahun 8 bulan.
- b. Cluster 1 memiliki jumlah data sebanyak 16 data dan mayoritas dari SMA yang berasal dari Asal 2, baik luar pulau jawa maupun dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 66,30, TOEFL sebesar 431,69, IPK sebesar 3,37 dengan lama studi 4 tahun 2 bulan.
- c. Cluster 2 memiliki jumlah data sebanyak 12 data dan mayoritas dari SMA yang berasal dari Asal 2, baik luar pulau jawa maupun dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 68,56, TOEFL sebesar 412,08, IPK sebesar 3,36 dengan lama studi kurang dari 4 tahun yaitu 3 tahun kurang 11 bulan.

d. Cluster 3 memiliki jumlah data sebanyak 27 data dan mayoritas berasal dari SMA, Asal 2 tetapi luar pulau jawa serta memiliki rata-rata nilai matematika sebesar 69,72, TOEFL sebesar 419,11, IPK sebesar 3,46 dengan lama studi 3 tahun 9 bulan.

Hasil dari evaluasi pola pada dataset Program Studi T dengan jumlah cluster sebanyak 4 cluster, maka diperoleh cluster 0 sebagai cluster yang terbaik untuk rekomendasi tepat waktu jika nilai rata-rata matematika besar sama dengan 73,89 dengan asal dari dalam pulau jawa dan rata – rata IPK sebesar 3,62.

4. Pada dataset Program Studi U dengan jumlah 90 data diperoleh hasil pengujian yang disajikan pada Tabel 12.

Tabel 12. Hasil Pengujian Dataset Program Studi U

Nilai K (Jumlah Cluster)	Hasil Silhouette Coefficient
2	0,65
3	0,44
4	0,23

Berdasarkan hasil pengujian pada dataset Program Studi U dapat ditarik kesimpulan bahwa jumlah cluster terbaik pada dataset Program Studi U yaitu 2. Sehingga diperoleh 2 cluster dengan evaluasi pola setiap cluster sebagai berikut :

- a. Cluster 0 memiliki jumlah data sebanyak 64 data dan mayoritas berasal dari SMA, Asal 2 dalam pulau jawa serta memiliki rata-rata nilai matematika sebesar 82,24, TOEFL sebesar 412,44, IPK sebesar 3,49 dengan lama studi 3 tahun 11 bulan.
- b. Cluster 1 memiliki jumlah data sebanyak 26 data dan mayoritas berasal dari SMA Asal 2 dalam pulau jawa, serta memiliki rata-rata nilai matematika sebesar 82,05, TOEFL sebesar 434,15, IPK sebesar 3,25 dengan lama studi lebih dari 4 tahun yaitu 4 tahun 8 bulan.

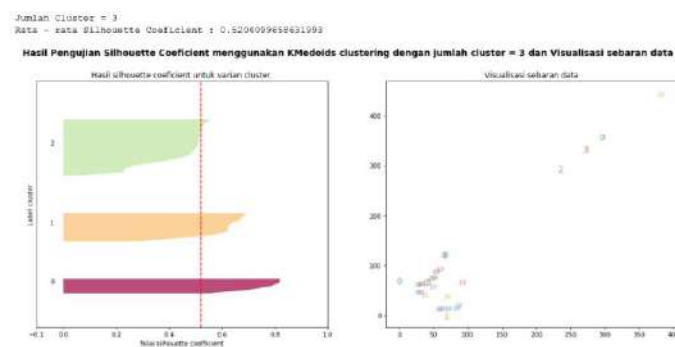
Hasil dari evaluasi pola pada dataset Program Studi U dengan 2 jumlah cluster, maka diperoleh cluster 0 sebagai cluster yang terbaik untuk rekomendasi tepat waktu jika nilai rata-rata matematika besar sama dengan 82,24 dengan asal dari dalam pulau jawa dan rata – rata IPK sebesar 3,49.

DISKUSI

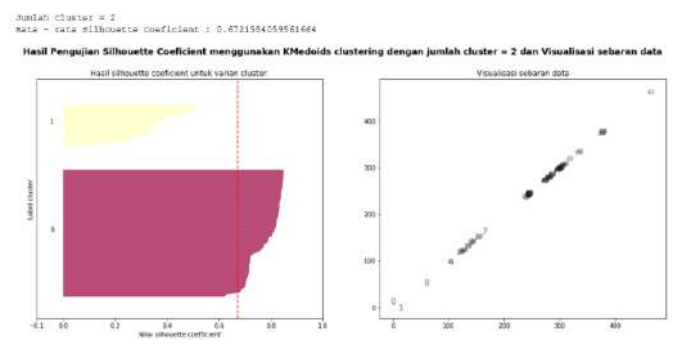
Penelitian dari dataset setiap prodi menghasilkan cluster-cluster yang direkomendasikan berdasarkan hasil pengujian data menggunakan metode *Silhouette Coefficient* dan visualisasi sebaran data yang telah diproses. Hasil dari pengujian tersebut memiliki range nilai dari -1 sampai 1. Jika hasil pengujian bernilai -1 atau mendekati nilai tersebut, maka dapat dikatakan bahwa nilai tersebut tidak bagus dan sebaliknya. Jika hasil pengujian bernilai 1 atau mendekati nilai maka dapat disimpulkan bahwa hasil pengujiannya tersebut bagus [12].

Gambar 3, 4, 5 dan 6 merupakan hasil pengujian yang dilakukan pada setiap dataset program studi, nilai diperoleh dari hasil pengujian diatas rata-rata yang diikuti dengan tingkat ketebalan plot siluet yang cukup besar. Sehingga dapat disimpulkan hasil pengujian yang dilakukan pada setiap program studi sudah bagus, karna dapat dilihat dari hasil pengujian bahwa semua hasil pengujian memiliki nilai yang mendekati 1 dan hanya pada data program studi T yang sedikit memiliki nilai negatif. Tingkat keberagaman data dan

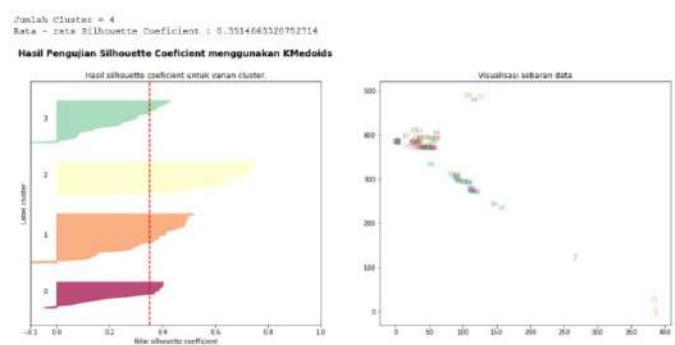
kemiripan data khususnya pada nilai atribut sangat mempengaruhi kedekatan data tersebut dalam sebuah *cluster*.



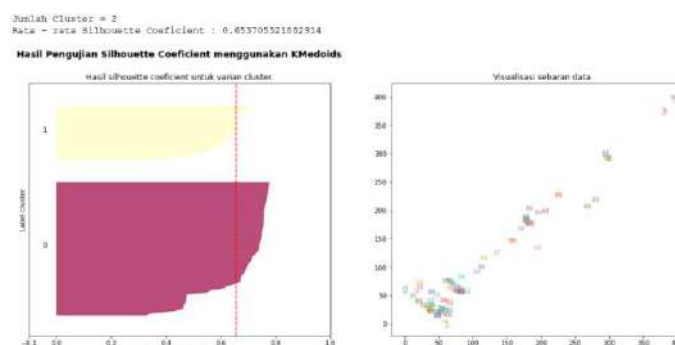
Gambar 3. Hasil Pengujian & Visualisasi Sebaran Data Program Studi R



Gambar 4. Hasil Pengujian & Visualisasi Sebaran Data Program Studi S



Gambar 5. Hasil Pengujian & Visualisasi Sebaran Data Program Studi U



Gambar 6. Hasil Pengujian & Visualisasi Sebaran Data Program Studi T

KESIMPULAN DAN SARAN

Berdasarkan hasil dari penelitian ini dapat ditarik kesimpulan, sebagai berikut:

- a. Penelitian menghasilkan beberapa data mahasiswa yang direkomendasikan dan semuanya berasal dari Pulau Jawa. Pada

dataset program studi R mahasiswa yang di rekomendasikan berjumlah 9 dengan *cluster* yang terbentuk sebanyak 3 *cluster*, dataset program studi S mahasiswa yang di rekomendasikan berjumlah 57 dengan *cluster* yang terbentuk sebanyak 2 *cluster*, dan dataset program studi U mahasiswa yang di rekomendasikan berjumlah 64 dengan *cluster* yang terbentuk sebanyak 2 *cluster*. Program studi R,S, dan U yang memiliki rata-rata nilai matematika minimal 82. Sedangkan untuk dataset program studi T diperoleh 35 data mahasiswa yang direkomendasikan dengan *cluster* yang terbentuk 4 *cluster* dan memiliki rata-rata nilai matematika sebesar 73,89.

- b. Hasil pengujian dengan menggunakan metode *Silhouette Coefficient* memperoleh hasil yang cukup bagus dengan nilai *silhouette coefficient* untuk dataset Program Studi R sebesar 0.52, untuk dataset Program Studi S sebesar 0.67, untuk dataset Program Studi T sebesar 0.35 dan untuk dataset Program Studi U sebesar 0.65.

Pada penelitian ini masih memiliki kekurangan, maka perlu adanya pengembangan untuk kedepannya untuk memperoleh hasil yang lebih bagus. Saran dari penulis yaitu perlu melakukan penambahan dataset dan atribut data, sehingga dapat meningkatkan akurasi dari pengelompokan dalam penelitian selanjutnya.

REFERENSI

- [1] W. Widarto, "Faktor Penghambat Studi Mahasiswa yang Tidak Lulus Tepat Waktu di Jurusan Pendidikan Teknik Mesin FT UNY," *J. Din. Vokasional Tek. Mesin*, vol. 2, no. 2, p. 127, 2017.
- [2] D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokan Wilayah Sebaran Cacat pada Anak," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 64, 2018.
- [3] D. F. Pramesti, M. Tanzil Furqon, and C. Dewi, "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 723-732, 2017.
- [4] L. Zahrotun, "Analisis Pengelompokan Jumlah Penumpang Bus Trans Jogja Menggunakan metode Clustering K-Means dan Agglomerative Hierarchical Clustering (AHC)," *J. Inform.*, vol. 9, no. 1, pp. 1039-1047, 2015.
- [5] L. Zahrotun, N. hutami Putri, and A. N. Khusna, "The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesisi Titles," in *12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2018.
- [6] L. Zahrotun, "Text Mining for Internship Titles Clustering Using Shared Nearest-Neighbor Method," *Comput. Eng. Appl.*, vol. 6, no. 3, 2017.
- [7] S. Defiyanti, M. Jajuli, and N. Rohmawati, "Optimalisasi K-MEDOID dalam Pengklasteran Mahasiswa Pelamar Beasiswa dengan CUBIC CLUSTERING CRITERION," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 1, pp. 211-218, 2017.
- [8] N. K. Kaur, U. Kaur, and D. Singh, "K-Medoid Clustering Algorithm- A Review," *Int. J. Comput. Appl. Technol.*, vol. 1, no. 1, pp. 42-45, 2014.
- [9] R. Handoyo, R. Rumani, and S. M. Nasution, "Perbandingan Metode Clustering Menggunakan Metode Single Linkage Dan K-Means Pada Pengelompokan

- Dokumen,” *JSM STMIK Mikroskil*, vol. 15, no. 2, pp. 73–82, 2014.
- [10] M. B. Al- Zoubi and M. A. Rawi, “An Efficient Approach for Computing Silhouette Coefficients,” *J. Comput. Sci.*, vol. 4, no. 3, pp. 252–255, 2008.
- [11] J. Han, Jiawei; Kamber, Micheline; Pei, *Data mining: Data mining concepts and techniques*. 2014.
- [12] W. Bagus, A. E. Budianto, and A. S. Wiguna, “Implementasi Metode K-Medoids Clustering untuk Mengetahui Pola Pemilihan Program Studi,” *J. Terap. Sains Teknol.*, vol. 1, no. 3, pp. 54–69, 2019.
- [13] G. S. Nugraha, Hairani, and R. F. P. Ardi, “Aplikasi Pemetaan Kualitas Pendidikan Di Indonesia Menggunakan Metode K-Means,” *J. MATRIK*, vol. 17, no. 2, pp. 13–23, 2018.

BIOGRAFI PENULIS



Herri Kurnia

Salah satu mahasiswa Universitas Ahmad Dahlan (UAD). Jurusan / Program Studi Teknik Informatika. Diterima menjadi mahasiswa Program Studi Teknik Informatika pada tahun 2016 di UAD.



Lisna Zahrotun, S.T., M.Cs.

Bekerja sebagai dosen Program Studi Teknik Informatika Fakultas Teknologi Industri di Universitas Ahmad Dahlan sampai sekarang. Bidang keahlian data mining, text mining, sistem pendukung keputusan.



Utaminingsih Linarti, S.T., M.T.

Bekerja sebagai dosen Program Studi Teknik Industri Fakultas Teknologi Industri di Universitas Ahmad Dahlan sampai sekarang. Bidang keahlian data rantai pasok, learning system.

TATA NAMA

Arti atau maksud dari symbol yang ada pada persamaan-persamaan yang digunakan.

Σ	berarti jumlah atau total keseluruhan
$ $	berarti Absolute atau mutlak
ϵ	berarti Elemen atau anggota
n	berarti jumlah data atau banyak data