Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 4 No. 3 (2020) 544 - 550 ISSN Media Elektronik: 2580-0760

Penentuan *Centroid* Awal *K-means* pada proses *Clustering* Data Evaluasi Pengajaran Dosen

Ridho Ananda[1], Achmad Zaki Yamani[2]
[1,2]Teknik Industri, Fakultas Rekayasa Industri dan Desain, Institut Teknologi Telkom Purwokerto
[1]ridho@ittelkom-pwt.ac.id, [2]zaki@ittelkom-pwt.ac.id

Abstract

Decision making about microteaching for lecturers in ITTP with the low teaching quality is only based on three lowest order from teaching values. Consequently, the decision is imprecise, because there is possibility that the lecturers are not three. To get the precise quantity, an analysis is needed to classify the lecturers based on their teaching values. Clustering is one of analyses that can be solution where the popular clustering algorithm is k-means. In the first step, the initial centroids are needed for k-means where they are often randomly determined. To get them, this paper would utilize some preprocessing, namely Silhouette Density Canopy (SDC), Density Canopy (DC), Silhouette (S), Elbow (E), and Bayesian Information Criterion (BIC). Then, the clustering results by using those preprocessing were compared to obtain the optimal clustering. The comparison showed that the optimal clustering had been given by k-means using Elbow where obtain four clusters and 0.6772 Silhouette index value in dataset used. The other results showed that k-means using Elbow was better than k-means without preprocessing where the odds were 0.75. Interpretation of the optimal clustering is that there are three lecturers with the lower teaching values, namely N16, N25, and N84.

Keywords: clustering, k-means, intial centroid, teaching, preprocessing

Abstrak

Pengambilan keputusan pelatihan *microteaching* dosen ITTP yang memiliki kualitas pengajaran rendah hanya didasarkan pada tiga urutan terendah dari nilai pengajaran dosen. Akibatnya, keputusan menjadi kurang akurat, dikarenakan ada kemungkinan dosen tersebut sejumlah lebih dari atau kurang dari tiga . Untuk mengetahui jumlah yang tepat, diperlukan suatu analisis untuk mengelompokkan kualitas pengajaran dosen di ITTP. Salah satu analisis yang dapat digunakan ialah analisis *clustering*. Algoritme *clustering* yang cukup popular ialah *k-means*. Dalam proses awal, *k-means* membutuhkan *centroid* awal yang seringkali ditentukan secara acak. Pada penelitian ini, akan digunakan beberapa *preprocessing* untuk menentukan *centroid* awal *k-means*, yakni *Silhouette Density Canopy* (SDC), *Density Canopy* (DC), *Silhouette* (S), *Elbow* (E), dan *Bayesian Information Criterion* (BIC). Hasil-hasil clustering dengan *preprocessing* tersebut selanjutnya dibandingkan untuk memperoleh *clustering* optimal. Hasil perbandingan menunjukan bahwa *k-means* menggunakan kriteria *Elbow* memberikan hasil *clustering* terbaik pada data nilai evaluasi dosen ITTP tahun 2018 dengan terbentuk empat *cluster* dan nilai index Silhouette 0.6772. Disimpulkan juga bahwa hasil *k-means* dengan kriteria *Elbow* memberikan hasil yang lebih baik dibandingkan *k-means* tanpa *preprocessing* dengan peluang 0.778 pada data yang digunakan. Interpretasi hasil *clustering k-means* dengan kriteria *Elbow* menunjukkan, ada 3 dosen dengan kinerja pengajaran yang rendah yaitu N16, N25, dan N84.

Kata kunci: *clustering, k-means, centroid awal*, pengajaran, *preprocessing*

1. Pendahuluan

Dosen merupakan salah satu komponen esensial dalam dunia pendidikan khususnya di perguruan tinggi. Tugas utama dosen ialah mentransformasikan, mengembangkan, dan menyebarluaskan ilmu pengetahuan, teknologi, dan seni. Hal tersebut tertuang dalam UU Nomor 14 Tahun 2005 tentang guru dan dosen. Tugas utama tersebut selanjutnya diuraikan dalam tiga kegiatan, yakni (1) pendidikan, (2) penelitian, dan (3) pengabdian masyarakat. Di dalam pendidikan, dosen memiliki tanggung jawab untuk mewujudkan visi misi suatu perguruan tinggi kemudian secara luas mewujudkan tujuan pendidikan nasional. Untuk mencapai tujuan tersebut, setiap perguruan tinggi memiliki aturan yang berkaitan dengan proses pendidikan dan pengajaran yang diatur dalam sistem

Diterima Redaksi: 18-04-2020 | Selesai Revisi: 06-06-2020 | Diterbitkan Online: 20-06-2020

penjamin mutu (SPM) perguruan tinggi berdasarkan UU Sisdiknas No. 20 tahun 2003 pasal 50 ayat 6.

IT Telkom Purwokerto (ITTP) merupakan institut di bawah pengelolaan *Telkom Foundation* yang memiliki 2 fakultas dengan total 8 program studi di dalamnya. Untuk memperoleh hasil yang berkualitas dalam proses pendidikan dan pengajaran oleh dosen, SPM di ITTP melakukan evaluasi terhadap pengajaran dosen secara periodik. Salah satu evaluasi tim SPM ITTP ialah mewajibkan mahasiswa mengisi kuesioner pengajaran dosen ketika menjelang ujian tengah semester (UTS) dan ujian akhir semester (UAS). Kuesioner tersebut terdiri dari beberapa pernyataan yang digeneralisasikan dalam tiga indikator utama yaitu sikap, sistematika, dan kesesuaian. Rata-rata dari ketiga indikator tersebut akan menjadi nilai evaluasi dari setiap dosen. Tiga dosen yang memperoleh nilai akhir terendah, selanjutnya akan ditugaskan untuk mengikuti *microteaching* pada waktu tertentu. Penentuan tiga dosen tersebut tentunya ada unsur subjektifitas, karena ada kemungkinan dosen yang masuk dalam kelompok nilai evaluasi terendah tidak hanya tiga namun bisa kurang atau bahkan lebih. Berdasarkan kemungkinan tersebut, maka akan lebih efektif apabila dalam penentuan kelompok dosen dengan nilai evaluasi terendah melibatkan suatu analisis berdasarkan data. Salah satu analisis yang dapat digunakan ialah analisis *clustering*.

Analisis *clustering* digunakan untuk mengelompokkan objek-objek menjadi beberapa kelompok berdasarkan indikator yang diamati, sehingga seluruh objek yang masuk dalam kelompok yang sama memiliki kemiripan yang tinggi, dibandingkan dengan objek yang masuk dalam kelompok berbeda. Salah satu algoritme *clustering* yang cukup populer ialah *k-means*. *K-means* telah diimplementasikan dalam beberapa bidang, diantaranya oleh Nurzahputra et. al. [1] dalam sistem pendukung keputusan pemilihan line-up pada pemain sepak bola. Oleh Cerah et. al [2] yang membandingkan *k-means* dengan region growing pada pengukuran luas wilayah hutan mangrove. Kemudian yang terbaru oleh Ananda dan Burhanuddin [3] *k-means* digunakan untuk menganalisis mutu pendidikan SMA program ilmu alam di Jawa Tengah.

Masalah yang mendasar dalam proses *k-means* ialah penentuan jumlah *cluster* secara optimal. Beberapa penelitian telah dilakukan dengan menggunakan algoritme tertentu atau kriteria tertentu untuk memperoleh jumlah *cluster* optimal. Kodinariya dan Makwana [4] mereviewer beberapa kriteria penentu banyaknya *cluster* optimal, yakni kriteria *Elbow*, *Bayesian Information Criterion* (BIC), dan *Silhouette*. Widiyanti [5] menentukan banyaknya *clustering* optimal dengan menggunakan pendekatan metode berhirarki dalam pengelompokkan provinsi di Indonesia. Penentuan banyaknya *clustering* optimal didasarkan pada kepadatan *clustering* juga telah dilakukan. Zhou et. al. [6] menggunakan algoritme *PD-Auto* untuk menentukan ambang batas *cluster* pada algoritme *Canopy* [7]. Yan et. al. [8] mengusulkan metode *Auto Density peak Detection* (ADPD) yang didasarkan pada konsep statistika untuk mengidentifikasi *centroid* pada dataset berdasarkan grafik keputusan. Penelitian berikutnya oleh Zhang et. al. [9] dengan mengusulkan algoritme *Density Canopy* (DC) dalam penentuan *centroid* dan jumlah *cluster* awal. DC merupakan algoritme penentu jumlah *cluster* yang *robust* terhadap *outlier* serta termasuk dalam algoritme deterministik. Kemudian Luka et. al [10] menggunakan kriteria *Silhouette* untuk memperoleh banyak *cluster* optimal pada algoritme *k-means*. Ananda [11] menambah kriteria *Silhouette* pada algoritme DC untuk menentukan jumlah *cluster* optimal. Penelitian terbaru oleh Ananda dan Burhanuddin [3] dengan menggunakan BIC untuk memperoleh banyak *cluster* optimal dalam menganalisis mutu pendidikan SMA program ilmu alam di Jawa Tengah.

Berdasarkan uraian tersebut maka dilakukan komparasi hasil *clustering* berdasarkan centroid awal yang diperoleh dari algoritme atau kriteria tertentu, yakni DC, SDC, serta metode *clustering* hirarki dengan kriteria BIC, *Elbow*, dan *Silhouette* untuk memperoleh hasil clustering yang optimal. Selanjutnya hasil optimal yang diperoleh, dibandingkan dengan algoritme *k-means* tanpa kriteria apapun untuk mengetahui apakah hasil optimal dari proses komparasi merupakan hasil optimal proses *clustering* secara umum berdasarkan data dan validasi clustering yang digunakan. Algoritme DC dan SDC dipilih dalam penentuan banyaknya *clustering* optimal berdasarkan kepadatan dikarenakan algoritme tersebut *robust* terhadap *outlier*. Hasil *clustering* terbaik selanjutnya dianalisis dengan menggunakan konsep *Gaussian* untuk mengetahui apakah hasil *clustering* merupakan nilai optimal dari algoritme *k-means* pada data yang digunakan. Hasil tersebut juga digunakan untuk sistem pengambilan keputusan terkait dosen yang akan diberikan pelatihan *microteaching* berdasarkan *cluster* dengan kemampuan terendah.

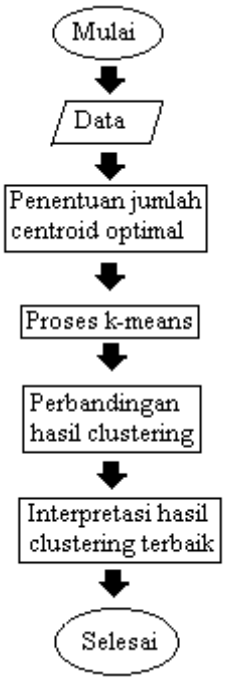
Paper ini disusun dengan urutan sebagai berikut: Bab II membahas tentang metode penelitian yang digunakan. Bab III membahas tentang hasil dan pembahasan. Pada bab terakhir diuraikan tentang kesimpulan dari hasil penelitian.

2. Metode Penelitian

2. 1. Alur penelitian

Alur penelitian pada paper ini memiliki 4 tahapan utama yakni tahap pertama penentuan jumlah *centroid* awal yang optimal, tahap kedua proses *clustering* dengan menggunakan algoritme *k-means*, tahap ketiga perbandingan hasil *clustering* menggunakan validasi *index Silhouette*, dan tahapan terakhir interpretasi hasil *clustering* terbaik. Pada tahap pertama, penentuan jumlah *centroid* awal menggunakan algoritme DC,

SDC, serta metode *clustering* hirarki dengan kriteria *Elbow*, BIC, dan *Silhouette*. Ilustrasi dari alur penelitian diberikan oleh Gambar 1.



Gambar 1. Alur Penelitian

2. 2. Data nilai evaluasi dosen ITTP tahun 2018

Data yang digunakan pada penelitian ini ialah data nilai evaluasi dosen tahun 2018 yang diperoleh dari SPM ITTP. Data tersebut berbentuk matriks dengan ordo 88×3 yang terdiri dari 88 dosen ITTP dan 3 indikator penilaian sebagai peubah yaitu sikap (X_1), sistematika pengajaran (X_2), dan kesesuaian pengajaran dengan literatur (X_3). Data berjenis interval dengan nilai antara 0 s.d. 100 untuk setiap indikator. Tabel 1 menampilkan beberapa contoh data yang digunakan pada penelitian ini.

Tabel 1. Beberapa contoh data nilai evaluasi dosen tahun 2018

Objek	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$
N1	76.75	79.97	76.21
N2	88.53	85.69	86.55
N3	85.66	84.04	86.27
N4	79.42	75.78	76.36
…	…	…	…

2. 3. Algoritme *Density Canopy K-Means*

Algoritme *Density Canopy* (DC) merupakan pengembangan dari algoritme *canopy* [7]. Algoritme ini telah diimplementasikan oleh Li *et. al* [12] dalam pengelompokkan komunitas tertentu dan Ananda *et al* [13] dalam sistem rekomendasi pemilihan peminatan. Langkah-langkah dari algoritme DC sebagai berikut:

Step 1. Menghitung rata-rata jarak antarobjek $(\bar{d}_E)$ dari dataset yang berbentuk matriks dengan ${}_n\mathbf{X}_p$ dengan n objek $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ dan p peubah dengan menggunakan persamaan 1.

$$\bar{d}_E = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_E(\mathbf{x}_i, \mathbf{x}_j) \qquad (1)$$

Step 2. Menentukan jumlah tetangga setiap objek-i atau $\rho(i)$ dengan persamaan 2, yaitu objek lain yang jaraknya terhadap objek-i kurang $\bar{d}_E$.

$$\rho(i) = \sum_{j=1}^{n} f\left(d_E(\mathbf{x}_i, \mathbf{x}_j) - \bar{d}_E\right) \qquad (2)$$

dengan ketentuan

$$f(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases}$$

Step 3. Memilih *centroid* pertama yaitu objek yang memiliki tetangga paling banyak. *Centroid* tersebut dan tetangganya dihapus dari dataset.

Step 4. Pada data sisa, dihitung $\rho(i)$ untuk setiap objek-i dengan prosedur di step 2.

Step 5. Untuk setiap objek-i, dihitung $a(i)$ yaitu jarak rata-rata objek-i dengan tetangganya. Dilakukan juga perhitungan $q(i)$ yaitu kepadatan lokal objek ke-i, dengan ketentuan:

$$q(i) = \begin{cases} min\, d_E(\mathbf{x}_i, \mathbf{x}_j) & \text{jika } \exists j \ni \rho(i) < \rho(j) \\ max\, d_E(\mathbf{x}_i, \mathbf{x}_j) & \text{jika } \forall j \ni \rho(i) > \rho(j) \end{cases}$$

Step 6. Menghitung bobot setiap objek-i yaitu $w(i)$, dengan persamaan 3.

$$w(i) = \rho(i) \times \frac{1}{a(i)} \times q(i). \qquad (3)$$

Step 7. Objek yang memiliki bobot terbesar dijadikan sebagai *centroid* berikutnya. *Centroid* tersebut dan tetangganya dihapus dari dataset.

Step 8. Proses kembali ke nomor 4 hingga tidak ada data yang tersisa.

Centroid yang diperoleh pada algoritme ini digunakan digunakan sebagai *centroid* awal untuk proses *K-means.*

2. 4. *Silhouette Density Canopy*

Algoritme *Silhouette Density Canopy* (SDC) merupakan pengembangan dari algoritme DC dengan *memasukkan* kriteria *Silhouette* [11]. Langkah-langkah dari SDC tidak jauh berbeda dengan algoritme DC, hanya sedikit perubahan pada pemilihan *centroid*. *Centroid* pertama diperoleh dengan menggunakan persamaan 4.

$$centroid \text{ pertama} = \underset{x_i}{argmax}[p(i) \times s(i)] \qquad (4)$$

dimana $s(i)$ merupakan nilai *Silhouette* dari objek ke-i dengan perhitungannya berdasarkan persamaan 5.

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

dengan $a(i)$ merupakan rata-rata jarak objek ke-i dengan tetanggannya dan $b(i)$ merupakan rata-rata jarak objek ke-i dengan seluruh objek yang bukan tetangganya.

Sedangkan *centroid* berikutnya ditentukan berdasarkan bobot $w(i)$ yang rumusnya dimodifikasi seperti pada persamaan 6.

$$w(i) = p(i) \times s(i) \times \frac{1}{a(i)} \times q(i) \quad (6)$$

dimana $p(i)$ merupakan tetangga dari objek ke-i. $s(i)$ merupakan nilai *Silhouette* dari objek ke-i. $a(i)$ merupakan jarak rata-rata objek ke-i dengan tetangganya. $q(i)$ kepadatan lokal objek ke-i.

2. 5. Algoritme *Clustering* Hirarki

Algoritme hirarki dalam *clustering* merupakan teknik penggabungan yang mana pada awalnya setiap objek merupakan satu *cluster* tersendiri. Lalu dua *cluster* terdekat digabungkan dan seterusnya sehingga diperoleh satu *cluster* yang beranggotakan seluruh objek. Dalam menggabungkan dua *cluster* digunakan ukuran kuantitatif berupa ukuran jarak antar *cluster* [14]. Algoritme hirarki yang digunakan pada paper ini ialah *single linkage* (SL), *average linkage* (AL), *complete linkage* (CL), dan *ward linkage* (WL). Algoritme hirarki tersebut hanya digunakan untuk menentukan *centroid* awal, yang kemudian akan dievaluasi dengan kriteria *Silhouette*, *Elbow*, dan BIC.

Kriteria *Silhouette* dihitung berdasarkan jarak kohesi dan pemisahan gerombol [11]. Formula pada kriteria *Silhouette* untuk setiap objek sebagaimana dalam persamaan 5 dimana $a(i)$ merupakan jarak rata-rata objek-i dengan seluruh objek pada *cluster* yang sama. $b(i)$ merupakan jarak rata-rata objek-i dengan seluruh objek pada *cluster* yang terdekat. Kriteria *Silhouette* diperoleh dari rata-rata seluruh $s(i)$. *Centroid* awal yang dipilih berdasarkan kriteria *Silhouette* ialah kriteria yang memenuhi persamaan 7.

$$S = max\, S(n, p) \quad (7)$$

dimana $S(n, p)$ merupakan hasil kriteria *Silhouette* pada algoritme hirarki p dengan *n cluster* yang terbentuk. Implementasi dari kriteria *Silhouette* untuk menentukan jumlah *cluster* terbaik juga telah dilakukan oleh Romanuke [15] untuk mengoptimalisasikan hasil *clustering*.

Kriteria *Elbow* menentukan *K cluster* optimal berdasarkan penambahan informasi yang paling maksimum, yang berarti apabila diubah menjadi *K*+1 *cluster* maka penambahan informasinya tidak begitu besar. Penambahan informasi tersebut digambarkan dengan jarak objek terhadap masing-masing *centroid* yang perhitungannya menggunakan persamaan 8.

$$\text{SSE} = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in S_k} \|\mathbf{x}_i - \mathbf{c}_k\|_2^2, \quad (8)$$

dengan *K* merupakan banyak *cluster*, S_k himpunan dari objek-objek yang masuk dalam *cluster k*, dan $\mathbf{c}_k$ merupakan *centroid* dari *cluster k*. Pemilihan *k cluster* dengan kriteria *Elbow* berdasarkan maksimum selisih nilai SSE pada pembentukan *k cluster* terhadap SSE dari pembentukan $k-1$ *cluster*. *Centroid* awal yang dipilih berdasarkan kriteria *Elbow* ialah kriteria yang memenuhi persamaan 9.

$$SSE = \max SSE(n, p), \quad (9)$$

dimana $SSE(n, p)$ merupakan hasil kriteria *Elbow* pada algoritme hirarki *p* dengan *n cluster* yang terbentuk. Penggunaan kriteria *Elbow* juga digunakan untuk menentukan menentukan jumlah *cluster* optimal pada klasifikasi yang menggunakan segmentasi gambar *colormap* [16].

Kriteria BIC yang digunakan pada penelitian ini menggunakan rumus pada persamaan 10.

$$\text{BIC} = n.\ln\left(\frac{\text{RSS}}{n}\right) + K * (p+1) * \ln(n), \quad (10)$$

dengan n merupakan banyak objek, K merupakan banyak *cluster*, dan RSS merupakan akar dari jumlah seluruh jarak objek dengan *centroid*nya. *Centroid* awal yang dipilih berdasarkan kriteria BIC ialah kriteria yang memenuhi persamaan 11.

$$BIC = \max BIC(n, p) \quad (11)$$

dimana $BIC(n, p)$ merupakan hasil kriteria BIC pada algoritme hirarki *p* dengan *n cluster* yang terbentuk.

2. 6. Validasi *Clustering*

Validasi dibutuhkan untuk mengetahui kualitas dari hasil *clustering*. Rendon *et. al* [17] menyampaikan bahwa validasi *clustering* terdiri dari validasi eksternal dan validasi internal. Pada paper ini digunakan validasi internal dikarenakan tidak adanya informasi *cluster* awal pada data, disamping itu validasi internal telah terbukti lebih baik dalam memvalidasi hasil *clustering* dibandingkan dengan validasi *eksternal* [18]. Validasi internal yang digunakan pada paper ini ialah index Silhouette. Validasi tersebut dipilih karena memiliki kinerja yang lebih baik dibandingkan dengan validasi internal yang lain [19]-[20]. Index *Silhouette* telah digunakan dalam penelitian di bidang pertanian [21] dan kesehatan[22]. Nilai *Silhouette* tiap objek diperoleh dari persamaan 5 dengan $a(i)$ merupakan jarak rata-rata objek-i dengan seluruh objek pada *cluster* yang sama. $b(i)$ merupakan jarak rata-rata objek-i dengan seluruh

objek pada *cluster* yang terdekat. Vendramin *et al* [23] menyatakan bahwa index *Silhouette* diperoleh dari rata-rata seluruh $s(i)$ dimana nilai tersebut berada pada interval -1 s.d. 1. Ia juga mengatakan bahwa semakin tinggi nilai $s(i)$ yang berarti mendekati 1, maka semakin tepat objek tersebut masuk dalam *cluster* berdasarkan nilai $s(i)$.objek yang memiliki $s(i) \approx 0$ mengartikan bahwa objek tersebut merupakan objek yang memiliki keunikan tinggi dibandingkan dengan cluster-cluster yang ada sehingga disarankan objek tersebut membentuk *cluster* tersendiri. Berdasarkan nilai-nilai $s(i)$ tersebut, maka apabila dalam perhitungan diperoleh $s(i) < 0$ mengindikasikan bahwa objek tersebut salah masuk *cluster* atau salah klasifikasi. Pada paper ini, juga akan diuji salah klasifikasi setiap objek pada hasil *clustering* berdasarkan nilai $s(i)$. Hasil *clustering* terbaik juga akan dibandingkan dengan *k-means* dengan menerapkan simulasi untuk mengetahui perbedaannya dengan nilai *k-means* tertinggi.

3. Hasil dan Pembahasan

3.1. Deskripsi data

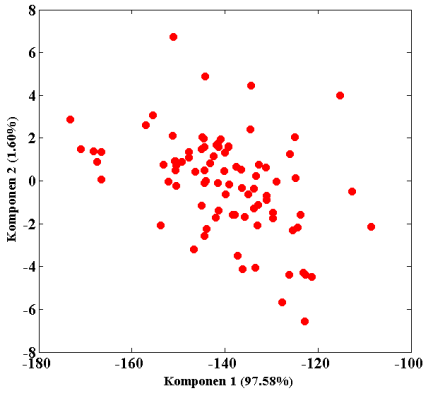
Deskripsi dari data nilai evaluasi dosen tahun 2018 disajikan pada Tabel 2. Pada tabel tersebut ditunjukkan peubah yang memiliki keragaman terbesar yakni X_2, disusul X_3, dan X_1. Rata-rata terbesar diperoleh peubah X_1, selanjutnya X_3 dan X_2. Selanjutnya, Tabel 3 memberikan ukuran korelasi antarpeubah. Pada tabel tersebut, diketahui bahwa korelasi antarpeubah pada X_1, X_2, dan X_3 bernilai lebih dari 0.900 sehingga ukuran tersebut menunjukkan adanya korelasi yang sangat kuat.

Tabel 2. Deskripsi data nilai evaluasi dosen

Peubah	μ	σ	Min	Maks
X_1	82.682	6.981	64.808	100
X_2	79.27	8.059	59.739	100
X_3	80.596	7.161	62.778	100

Tabel 3. Korelasi antarpeubah

	X1	X2	X3
X1	1		
X2	0.922	1	
X3	0.952	0.969	1



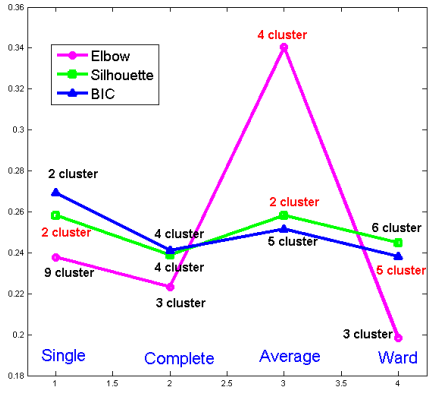
Gambar 2. Sebaran objek data dalam 2D

Selanjutnya Gambar 2 memberikan visualisasi dari sebaran data nilai evaluasi dosen tahun 2019. Sebaran tersebut diperoleh dengan menggunakan *Principal Component Analysis* (PCA) dengan ukuran jarak tiap objek menggunakan jarak Euclidean.

3.2. Hasil penentuan *centroid* awal optimal

Setelah melihat sebaran data untuk masing-masing peubah serta visualisasi dalam dimensi 2, selanjutnya dilakukan proses preprocessing untuk menentukan jumlah *centroid* yang optimal pada data nilai evaluasi dosen ITTP tahun 2018. Pada algoritme DC diperoleh 4 *centroid*. Pada algoritme SDC juga diperoleh 4 *centroid* namun dengan letak yang berbeda.

Pada metode penelitian telah dibahas bahwa kriteria *Silhouette* optimal apabila nilainya maksimum, kriteria *Elbow* optimal apabila nilainya maksimum, dan kriteria BIC optimal apabila nilainya minimum. Pada Gambar 3 diperoleh jumlah *centroid* dengan kriteria *Silhouette* saat menggunakan metode hirarki SL atau AL dengan jumlah 2 *centroid* . Sedangkan dengan kriteria *Elbow* diperoleh jumlah *centroid* optimal sebanyak 4 dengan menggunakan metode hirarki AL. Selanjutnya pada kriteria BIC, jumlah *centroid* optimal yang diperoleh sebanyak 5 dengan menggunakan metode WL. Hasil penentuan jumlah *centroid* optimal dengan metode DC dan SDC serta kriteria *Silhouette*, *Elbow*, dan BIC digunakan sebagai informasi awal untuk proses *k-means*.



Gambar 3. Grafik hasil kriteria tertentu pada metode hirarki

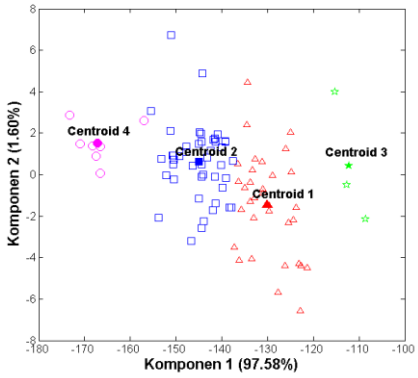
3.3. Perbandingan hasil *clustering*

Hasil proses *k-means* dengan jumlah *centroid* awal berdasarkan kriteria atau algoritme tertentu pada pada data nilai evaluasi dosen ITTP tahun 2018 selanjutnya dibandingkan dengan memperhatikan nilai Index *Silhouette* (IS) dan salah klasifikasi (SK).

Tabel 4. Validasi hasil *clustering k-means*

Ket	DC	SDC	*S*	*E*	BIC
nC	4	4	2	4	5
IS	0.6099	0.6099	0.6707	0.6772	0.658
SK	5	5	9	0	2
$\%SK$	5.68	5.68	10.23	0	2.27

Tabel 4 memberikan informasi hasil *clustering k-means* dengan *Density Canopy* (DC), *Silhouette Density Canopy* (SDC), *Silhouette* (S), *Elbow* (E), dan *Bayesian information criterion* (BIC). Notasi nC merupakan jumlah *cluster*. Notasi *IS* merupakan nilai dari index Silhouette. Notasi *SK* merupakan banyaknya objek yang salah klasifikasi. Dan notasi $\%SK$ merupakan perbandingan banyaknya objek yang salah klasifikasi dengan seluruh objek. Pada tabel tersebut, apabila dicermati secara komprehensif maka hasil *clustering* yang terbaik ialah *clustering k-means* dengan kriteria *Elbow*. Hal itu dikarenakan nilai index *Silhouette* yang paling besar serta salah klasifikasi yang paling minimum bahkan tidak ada objek yang salah klasifikasi. Gambar 4 memberikan visualisasi hasil *clustering k-means* dengan kriteria *Elbow*.



Gambar 4. Hasil clustering *k-means* dengan kriteria *Elbow*.

Selanjutnya hasil *clustering k-means* dengan kriteria *Elbow* akan dibandingkan dengan algoritme *k-means* tanpa kriteria apapun. Hal ini dilakukan untuk mengetahui apakah hasil *clustering* dengan penambahan kriteria *Elbow* merupakan nilai optimal dari *k-means* atau ada nilai yang lebih tinggi lagi berdasarkan *index Silhouette*. Untuk memperoleh jawaban tersebut, dilakukan simulasi *k-means* dengan perulangan sebanyak 1000 kali pada jumlah *cluster* dari 3 s.d. 11. Untuk *k-means* dengan kriteria *Elbow* tidak dilakukan perulangan karena merupakan model deterministik yang memberikan hasil konsisten.

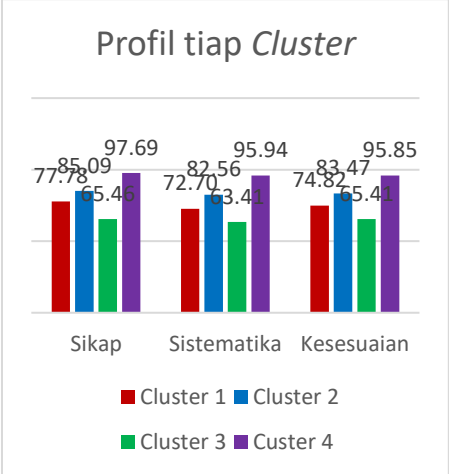
Tabel 5. Hasil simulasi *k-means*

Cluster	*IS*	$k > ke$	*n*
3	0.6416	NaN	0
4	0.6466	0.6807	291
5	0.6451	NaN	0
6	0.6537	NaN	0
7	0.6183	0.6794	222
8	0.574	NaN	0
9	0.5452	NaN	0
10	0.529	NaN	0
11	0.5163	NaN	0

Tabel tersebut memberikan informasi *IS* yaitu rata-rata *index Silhouette* secara keseluruhan pada jumlah *cluster* tertentu. Notasi $k > ke$ merupakan rata-rata hasil *IS k-means* yang lebih baik daripada *k-means* dengan *Elbow*. Notasi *n* merupakan banyaknya perulangan yang terjadi untuk *k-means* lebih baik daripada *k-means* dengan kriteria *Elbow*. NaN (*Not available number*) mengindikasikan tidak adanya nilai *index Silhouette* pada *k-means* yang melebihi *k-means* dengan kriteria *Elbow*. Hasil *IS* pada tabel menunjukkan bahwa hasil *k-means* dengan kriteria *Elbow* lebih baik daripada rata-rata hasil *clustering k-means*. Pada jumlah *cluster* tertentu yaitu 4 *cluster* dan 7 *cluster*, pada beberapa perulangan, *k-means* tanpa *Elbow* menunjukkan hasil yang lebih baik daripada dengan *Elbow*. Pada 4 *cluster*, kualitas *cluster k-means* yang lebih baik tersebut hanya terjadi sebanyak 291 dari 1000 perulangan sehingga apabila menggunakan konsep peluang Gaussian, kemungkinan mendapatkan *cluster* dengan kualitas tersebut hanya sebesar 0.291. Begitu pula pada 7 *cluster*, hasil yang lebih baik hanya diperoleh sebanyak 222 dari 1000 perulangan, sehingga peluang memperoleh *cluster* tersebut hanya 0.222. Nilai peluang dari *k-means* tanpa kriteria *Elbow* untuk memperoleh *cluster* yang lebih baik dari pada *k-means* disertai *Elbow* relatif kecil sehingga akan berisiko mendapatkan kualitas *cluster* yang buruk apabila hanya menggunakan *k-means* tanpa kriteria *Elbow*. Dengan kata lain, hasil *k-means* dengan kriteria *Elbow* akan menghasilkan kualitas yang lebih baik daripada *k-means* tanpa kriteria apapun dengan peluang 0.778. Hasil simulasi tersebut juga menunjukkan bahwa *kmeans* dengan kriteria *Elbow* pada data kinerja dosen cukup bagus walaupun belum mencapai hasil optimal dari *k-means*.

3.4. Interpretasi Data

Hasil perbandingan *k-means* dengan algoritme *preprocessing* atau kriteria tertentu menunjukkan bahwa *k-means* dengan kriteria *Elbow* memberikan hasil yang lebih baik dibandingkan dengan *preprocessing* atau kriteria lain yang digunakan pada penelitian ini untuk data nilai evaluasi dosen tahun 2018.



Gambar 5. Profil tiap *cluster* data nilai evaluasi dosen tahun 2018

Berdasarkan hasil tersebut, maka *k-means* dengan kriteria *Elbow* digunakan untuk menginterpretasi kinerja

pengajaran dosen. *Clustering* dari algoritme *k-means* dengan kriteria *Elbow* menghasilkan empat *cluster* dengan sebaran 33 (37.5%) dosen di *cluster* 1, 45 (51.14%) dosen di *cluster* 2, 3 (3.14%) dosen di *cluster 3*, dan 7(7.95%) dosen di *cluster* 4. Profil dari masing-masing *cluster* ditunjukkan pada Gambar 5. Pada Gambar tersebut, secara intuitif diperoleh informasi bahwa *cluster* 3 memiliki kualitas yang rendah untuk setiap indikator penilaian (sikap, sistematika, dan kesesuaian). Dosen yang masuk dalam *cluster* 3 tersebut ialah N16, N25, dan N84.

4. Kesimpulan

Berdasarkan hasil dan pembahasan yang telah diperoleh, disimpulkan bahwa *k-means* dengan kriteria *Elbow* menghasilkan nilai yang paling optimal dibandingkan dengan algoritme DC dan SDC atau kriteria *Silhouette* dan *BIC* pada data nilai evaluasi dosen ITTP tahun 2018. Selanjutnya, hasil perbandingan *k-means* menggunakan kriteria *Elbow* dengan *k-means* sendiri menunjukkan bahwa hasil *k-means* dengan kriteria *Elbow* belum memberikan hasil yang optimal dari *k-means* itu sendiri pada data evaluasi pengajaran dosen tahun 2018. Hanya saja, apabila hanya menggunakan *k-means* tanpa kriteria *Elbow*, maka peluang untuk mendapatkan hasil *clustering* yang lebih baik relatif kecil. Interpretasi hasil *clustering* pada data nilai evaluasi dosen ITTP tahun 2018 ialah ada 3 dosen yaitu dosen dengan initial N16, N25, dan N84 yang masuk dalam *cluster* dengan kemampuan terendah pada tiga indikator penilaian. Berdasarkan kondisi itu, maka ketiga dosen tersebut dapat disarankan untuk mengikuti pelatihan *microteaching*. Untuk penelitian selanjutnya, dapat melakukan modifikasi pada kriteria *Elbow* agar diperoleh hasil optimal dari *k-means* berdasarkan ukuran *Silhouette.*

Ucapan Terimakasih

Ucapan terima kasih kami sampaikan kepada unit SPM ITTP yang telah menyediakan data sehingga penelitian ini dapat berjalan dengan lancar.

Daftar Rujukan

[1] Nurzahputra A., Pratana A. R., Puwinarko A., 2017. Sistem Pendukung Keputusan Pemilihan *line-up* Pemain Sepak Bola Menggunakan Metode Fuzzy Multiple Attribute Decision Making dan K-means Clustering. *Jtsiskom*, 5(3),pp. 106-109.

[2] Cerah T. P. N., Nurhayati O. D., Isnanto R. R., 2019. Perbandingan Metode Segmentasi K-means Clustering dan Segmentasi Region Growing untuk Pengukuran Luas Wilayah Hutan Mangrove. *Jtsiskom*, 7(1),pp. 31-37.

[3] Ananda R., Burhanuddin A., 2019. Analisis Mutu Pendidikan Sekolah Menengah Atas Program Ilmu Alam di Jawa Tengah dengan Algoritme K-means Terorganisir. *Inista*, 2(1), pp. 065-072.

[4] Kodinariya T. M., Makwana P. R., 2013. Reviewer on Determining Number of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6), pp. 90-95.

[5] Widiyanti, W., 2017. *Sentroid awal metode k-means dengan pendekatan metode berhirarki dalam pengelompokkan provinsi di Indonesia*. Skripsi, Indonesia: Institut Pertanian Bogor.

[6] Zhou R., Zhang S., Chen C., Ning L., Zhang Y., 2016. A distance and density-based clustering algorithm using automatic peak detection, *2016 IEEE International Conference on Smart Cloud*. New York, 18-20 Nov 2016, IEEE.

[7] Kumar A., Ingle Y. S., Pande A., Dhule P., 2014. Canopy Clustering: A Review on Pre-clustering Approach to K-means Clustering. *IJIACS*, 3(5), pp. 22-29.

[8] Yan H., Lu Y., Ma H., 2018. Density-based Clustering Using Automatic Density Peak Detection, *In Proceedings of the 7th International Conference on Pattern Recognition Applicaions and Methods (ICPRAM 2018)*, China.

[9] Zhang G., Zhang C., Zhang H., 2018. Improved K-means Algorithm Based on Density Canopy. *Journal Knowledge-based Systems*, 145, pp. 289-297.

[10] Luka S. S. P., Candiasa I. M., Aryanto K. Y. E., 2019. Analisis Pembentukan Kelompok Diskusi Panel Siswa Menggunakan Algoritma Fuzzy C-means dan K-means. *Jurnal Pendidikan Teknologi dan Kejuruan*, 16(2), pp. 267-277.

[11] Ananda R., 2019. Silhouette Density Canopy K-means for Mapping the Quality of Education Based on the Results of the 2019 National Exam in Banyumas Regency. *Jurnal Khazanah Informatika*, 5(2), pp. 158-168.

[12] Li M., Wang H., Long H., Xiang J., Wang B., Xu J., Yang J., 2019. Community Detection and Visualization in Complex Network by the Density-Canopy-Kmeans Algorithm and MDS Embedding, *IEEE Access*, 20 August 2019, IEEE.

[13] Ananda R., Naf'an M. Z., Arifa A. B., Burhanuddin A., 2020. Sistem Rekomendasi Pemilihan Peminatan Menggunakan Density Canopy K-means. *Jurnal Resti*, 4(1), pp. 172-179.

[14] Johnson R. A., Wichern D. W., 2007. *Applied Multivariate Statistical Analysis*. 6th ed. New York (US): Pearson Education.

[15] Romanuke V. V., 2018. Optimization of A Dataset For A Machine Learning Task by Clustering and Selecting Closest-To-The-Centroid Objects,. *Вісник Хмельницького національного університету*, 1(6), pp. 263-265.

[16] Ali N., Hamilton N., Calaf M., Cal R. B., 2019. Classification of the Reynolds Stress Anisotropy Tensor in Very Large Thermally stratified Wind Farms Using Colormap Image Segmentation. *Journal of Renewable and Sustainable Energy*, 11, pp. 1-9.

[17] Rendon E., Abundez I. M., Gutierrez C., Zagal S. D., Arizmendi A., Quiroz E.M., Arzate H.E., 2011. A Comparison of Internal and External Cluster Validation Indexes, *in Proceeding of the 2011 American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications*. Amerika, January 2011.

[18] Rendon E., Abudez I., Arizmendi A., Quiroz E. M., 2011. Internal Versus External Cluster Validation Indexes. *International Journal of Computers and Communications*, 5(1), pp. 27-34.

[19] Khairati A. F., Adlina A. A., Hertono G. F., Handari B. D., 2019. Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA, *Prosiding Seminar Nasional Matematika (PRISMA)*. Semarang, February 2018.

[20] Baarsch J., Celebi M. E., 2012. Investigation of Internal Validity Measures for K-means Clustering, *In Proceedings of the IMECS*. March 14 – 16, 2012, Hong Kong.

[21] Munthe A. D., 2019. Penerapan Clustering Time Series untuk Menggerombolkan Provinsi di Indonesia Berdasarkan Nilai Produksi Padi. *Jurnal Litbang Sukowati*, 2(2), pp. 1-11.

[22] Suprihatin, Utami Y. R. W., Nugroho D., 2019. K-Means Clustering untuk Pemetaan Daerah Rawan Demam Berdarah. *Jurnal TIKomSIN*, 7(1), pp. 8-16, 2019.

[23] Vendramin L., Campello R. J. G. B., Hruschka E. R., 2009. On the Comparison of Relative Clustering Validity Criteria, *in Proceedings of the 2009 SIAM International Conference on Data Mining*. April 30 – May 2, 2009, Nevada, USA