Pemetaan Karakteristik Mahasiswa Penerima Kartu Indonesia Pintar Kuliah (KIP-K) menggunakan Algoritma K-Means++

Fitri Nuraeni[1]*, Dede Kurniadi[2], Gisna Fauzian Dermawan[3]
Teknik Informatika, Jurusan Ilmu Komputer
Institut Teknologi Garut
Garut, Indonesia
fitri.nuraeni@itg.ac.id [1], dede.kurniadi@itg.ac.id [2], 1806122@itg.ac.id [3]

Abstract*— Knowledge about mapping the characteristics of students receiving KIP-K in tertiary institutions can use data mining, namely the clustering technique. The mapping of these characteristics is carried out by grouping students based on academic and non-academic attributes using the K-Means++ algorithm, which can reduce the number of repetitions in the data grouping process. This study uses the Cross-Industry Standard Process for Data Mining (CRIPS-DM) method and the clustering algorithm, namely k-means++. This study produced a clustering model with a k = 2 based on the elbow method graph with the most significant silhouette coefficient value of 0.7523 and the smallest davies bouldine index (DBI) of 0.49053. From mapping the characteristics of KIP-K recipient students, knowledge is obtained that can be used as material for decision-making in tertiary institutions in selecting KIP-K applicants to minimize academic problems for KIP-K recipient students in the future.

***Keywords*— clustering, DBI, elbow method, k-means++, silhouette coefficient**

***Abstrak*— Pengetahuan baru mengenai pemetaan karakteristik mahasiswa penerima KIP-K pada perguruan tinggi dapat menggunakan penggalian data yaitu teknik *clustering*. Pemetaan karakteristik ini dilakukan dari hasil pengelompokan mahasiswa berdasarkan atribut akademik dan non-akademik menggunakan algoritma *K-Means++* yang dapat menurunkan jumlah perulangan dalam proses pengelompokan datanya. Dengan menggunakan metode *Cross-Industry Standard Process for Data Mining* (CRIPS-DM) dan algoritma *clustering* yaitu *k-means++*. Dari penelitian ini, dihasilkan model *clustering* dengan nilai *k*=2 berdasarkan grafik metode *elbow* dengan nilai *silhouette coefficient* terbesar yaitu 0.7523 dan *davies bouldine index* (DBI) terkecil yaitu 0.49053. Dari hasil pemetaan karakteristik mahasiswa penerima KIP-K ini, didapatkan pengetahuan yang dapat menjadi bahan pengambilan keputusan perguruan tinggi penyelenggaran dalam penyeleksian pendaftar KIP-K sehingga meminimalisir masalah akademik mahasiswa penerima KIP-K di kemudian hari.**

***Kata Kunci*— klasterisasi, DBI, k-means++, kip-k, metode elbow, silhouette coefficient**

I. PENDAHULUAN

Penyeleksian calon penerima KIP-K masih belum efektif, terlihat dari munculnya permasalahan mahasiswa penerima KIP-K selama proses pembelajaran. Permasalahan tersebut diantaranya nilai akademik yang dibawah rata-rata, kurangnya aktifitas mahasiswa dalam kegiatan akademik dan non-akademik di kampus, serta kurangnya motivasi dalam belajar sehingga mengundurkan diri sebagai mahasiswa. Untuk meminimalisir kemunculan masalah tersebut, maka perlu adanya pengetahuan baru mengenai pemetaan mahasiswa penerima KIP-K berdasarkan kesamaan karakter masing-masing. Karakter mahasiswa dapat diidentifikasi dari aspek akademik seperti Indeks Prestasi (IP) dan aspek non-akademik seperti faktor orang tua dan lingkungan tempat tinggal[1][2][3]. Selanjutnya, pemetaan karakter mahasiswa ini dapat digunakan sebagai bahan pengambilan kebijakan perguruan tinggi[4], salah satunya kebijakan seleksi penerima KIP-K.

Beberapa perguruan tinggi sudah menggunakan pemetaan karakter mahasiswa sebagai bahan pengambilan kebijakan. Tahun 2018 terdapat penelitian mengenai klasifikasi mahasiswa yang berpotensi mendapat beasiswa di perguruan tinggi menggunakan algoritma *k-Nearest Neighbor* (k-NN) dengan atribut nilai IPK dan status ekonomi yang dimiliki mahasiswa[5]. Satu tahun selanjutnya, pengelompokan mahasiswa penerima beasiswa KIP-K telah dilakukan menggunakan teknik *clustering* dengan tiga *cluster*, yaitu kategori berhak menerima, dengan pertimbangan, dan tidak berhak menerima, dilihat dari kriteria ekonomi dan prestasi mahasiswa di kampus[6]. Penelitian serupa mengenai pengelompokan penerima beasiswa dengan atribut nilai raport, prestasi disekolah, hasil test dan juga tingkat ekonomi keluarga [7]. Disamping itu, telah dilakukan pula pemetaan karakter mahasiswa untuk analisis pola masa studi dengan menggunakan data akademik dan biodata mahasiswa[8]. Pada tahun 2021 pengelompokan penerimaan beasiswa Unit Pengumpulan Zakat (UPZ) digunakan untuk menentukan mahasiswa yang layak, tidak layak, dan dipertimbangkan untuk mendapatkan bantuan beasiswa tersebut[9]. Selanjutnya, pada tahun 2022 hasil pengelompokan dengan kemiripan data non-akademik digunakan sebagai pemetaan potensi calon mahasiswa pada sebuah perguruan tinggi[10].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i3.1439, Copyright ©2022
Submitted : July 19, 2022; Revised : January 7, 2023; Accepted : January 9, 2023; Published : January 25, 2023

Berdasarkan penelitian-penelitian tersebut, diketahui bahwa algoritma *K-Means* dan *K-Medoids* menjadi algoritma yang dapat digunakan dalam proses pengelompokan tanpa supervisi[11] dan hanya berdasarkan kemiripan data (*clustering*) untuk mendapatkan pemetaan karakter mahasiswa. Perbandingan dari kedua algoritma tersebut menghasilkan nilai akurasi *K-Means* lebih tinggi dibanding *K-Medoids*[12]. K-means memiliki langkah awal dengan menentukan nilai dari *k* yang akan menentukan jumlah klaster yang dibangun. Untuk mendapatkan nilai *k* yang optimal, maka digunakan beberapa metode untuk mengevaluasi proses klasterisasi diantaranya yaitu perbandingan *silhouette coefficient*[13], evaluasi *Davies-Bouldin Indeks* (DBI)[14], metode *elbow*[15][16].

Disamping itu, algoritma *K-Means* ini memiliki beberapa cara dalam menentukan inisial centroid pada perulangan pertama. Cara yang populer adalah dengan teknik acak, yaitu mengambil sembarang titik data sebagai centroid setiap klaster[17]. Teknik acak ini dapat berakibat pada jumlah perulangan pengelompokan data yang meningkat untuk mencapai kondisi konvergen[18]. Oleh karena itu, mulai digunakan cara lain untuk menentukan titik centroid awal, salah satunya yaitu *K-means++* yang mengambil titik centroid yang saling berjauhan satu dengan lainnya[19].

Memperhatikan penelitian sebelumnya, maka penelitian ini bertujuan untuk mendapatkan pemetaan karakteristik mahasiswa penerima KIP-K berdasarkan klaterisasi dengan atribut ekonomi mahasiswa dan juga nilai *Index* Prestasi (IP) yang mereka miliki, serta menambahkan jarak tempat tinggal ke kampus sebagai atribut baru. Algoritma yang digunakan adalah *K-means* dan *K-means++* untuk mengetahui penurunan jumlah perulangan pengelompokan data. Selain itu, penelitian ini membandingkan hasil klaterisasi menggunakan metode evaluasi DBI, *silhouette coefficient* dan metode *elbow*, untuk mendapatkan nilai *k* yang optimal. Hasil klasterisasi ini disajikan dalam pemetaan karakteristik mahasiswa penerima KIP-K yang selanjutnya dapat dijadikan bahan pembuat kebijakan seleksi dan skala prioritas penentuan penerima KIP-K periode yang akan datang.

II. METODE PENELITIAN

A. Tahapan Penelitian

Penelitian ini menggunakan metode Cross Industry Standard Process for Data Mining (CRIPS-DM) yang memiliki enam tahapan mulai pemahamaan bisnis, pemahaman data, persiapan data, pemodelan, evaluasi dan tahapan terakhir penyebaran hasil *data mining*. Pada Fig 1., tahap awal pemahaman bisnis didapatkan masalah yang dibahas yaitu mengenai karakteristik mahasiswa penerima KIP-K dalam proses pembelajarannya yang tidak sesuai harapan pengelola perguruan tinggi, karena kinerja akademik dan non-akademik yang tidak sesuai harapan. Sehingga tujuan dari proses penggalian data ini adalah melakukan pemetaan karakteristik mahasiswa KIP-K dengan teknik klasterisasi.

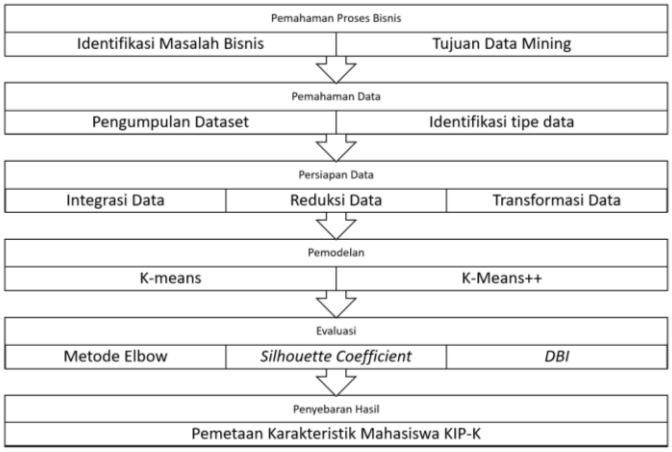

Fig. 1. Tahapan penelitian

Tahapan kedua pemahaman data, dilakukan pengumpulan data yang selanjutnya dilakukan identifikasi dan evaluasi kualitas data, untuk menyesuaikan bentuk data dengan pemodelan yang akan dibuat. Dataset yang digunakan berisi 225 data mahasiswa yang diambil dari penerima KIP-K tahun 2014 sampai 2019 dari perguruan tinggi swasta di daerah Kabupaten Garut. Dataset ini menggabungkan data pendaftaran mahasiswa KIP-K dan data akademik mahasiswa KIP-K selama 4 semester pertama seperti pada Fig 3 dibawah ini.

NO	NPM	JURUSAN	Pekerjaan	Jumlah Tanggungan	Penghasilan	Prestasi	Alamat	Ip Sem 1	Ip Sem 2	Ip Sem 3	Ip Sem 4
1	1903013	Teknik Industri	3	5	3,600,000	0	Kp. Patrol RT.03/RW.09 Ds. Sukaratu Kec. Banyuresmi Kab. Garut	4	3,89	3,71	3,92
2	1903014	Teknik Industri	6	5	12,000,000	0	Kp. Cikoleberes Ds. Tanjung Karang Kec. Cigalontang Kab. Garut	3,63	3,83	3,52	3,58
3	1903015	Teknik Industri	6	3	12,000,000	0	Kp. Sekbrak RT. 02/ RW. 03 Ds. Mekarbakti Kec. Bungbulang Kab. Garut	3,84	3,72	3,43	3,5
4	1903016	Teknik Industri	6	4	12,000,000	0	Kp. Tegalpanjang RT. 01/ RW. 16 Ds. Tegalpanjang Kec. Sucinaraja	3,89	3,89	3,52	3,67
5	1911015	Teknik Sipil	6	4	24,000,000	0		3,11	3,42	2,18	2,95
6	1911035	Teknik Sipil	3	6	18,000,000	0	Kp. Cikenceh 1 Ds. Harumansari Kec. Kadungora Kab. Garut	3,79	3,68	3,68	3,84
7	1911017	Teknik Sipil	5	4	12,000,000	0	Kp. Pataruman RT. 01/ RW. 12 Ds. Pataruman Kec. Tarogong Kidul	3,79	3,53	3,47	3,5
8	1906029	Teknik Informatika	5	7	12,000,000	0	Kp. Kalapa Sewu Rt. 01/ RW. 08 Ds. Sinar Jaya Kec. Bungbulang	3,67	4	3,87	3,85
9	1906030	Teknik Informatika	6	4	18,000,000	0	Kp. Dungus Maung RT.01/RW.04 Ds. Sirna Galih Kec. Cisurupan Kab. Garut	3,33	3,59	3,52	3,2
10	1906031	Teknik Informatika	5	4	12,000,000	0	Jl. Ibu Noch Kartanegara RT. 1/ RW. 19 Kp. Bentarhilir Kel. Sukamentri Kab. Garut	4	4	3,74	3,85

Fig. 2. Tampilan Data Penerima KIP-K

Tahapan selanjutnya adalah persiapan data, dimana dataset pada Fig.3 diatas dilakukan reduksi dimensi data sehingga atribut yang digunakan yaitu atribut jenis kelamin (JK), pekerjaan orang tua (PK), penghasilan orang tua (PH), jumlah tanggungan orang tua (JT), jumlah prestasi (PR), jarak dari rumah ke kampus (JR), IP semester satu (1), IP semester dua (2), IP semester tiga (3) dan IP semester empat (4), seperti yang tersaji dalam Table I.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i3.1439, Copyright ©2022
Submitted : July 19, 2022; Revised : January 7, 2023; Accepted : January 9, 2023; Published : January 25, 2023

TABLE I. DATA PENERIMA KIP-K

NPM	JK	PK	JT	PH	PR	JR	1	2	3	4
140336	2	2	6	4	4	8.6	4,00	3,91	3,82	3,78
141149	1	2	6	6	2	3.1	3,72	3,49	3,45	3,49
140304	1	5	7	6	4	76	3,32	3,15	3,16	3,14
140314	1	3	6	8	0	19	2,63	1,92	2,30	2,50
140012	2	5	4	3	2	13	3,63	3,38	3,43	3,44
141060	2	3	5	6	0	6	3,78	3,72	3,52	3,58
140015	1	6	4	10	0	6	2,74	2,62	2,80	2,83
...	...	...	...	...	...	...	...	...	...	...
190037	1	6	4	8	0	73	3,44	3,87	3,82	3,00

Tabel I merupakan hasil transformasi data yang semula terdapat bentuk polinominal menjadi numerik. Penentuan kategori atribut menggunakan aturan berikut:

1. Atribut JK terbagi menjadi kategori: (1) Laki-Laki (2) Perempuan;
2. Atribut PK terbagi menjadi kategori: (1) PNS; (2) Pegawai Swasta; (3) Wirausaha; (4) TNI/Polri; (5) Petani, Nelayan; (6) Lainnya;
3. Atribut PH terbagi menjadi kategori: (1) <250.000; (2) 250.001 – 500.000; (3) 500.001 - 750.000; (4) 750.001 - 1.000.000; (5) 1.000.001 - 1.250.000; (6) 1.250.001 - 1.500.000; (7) 1.500.001 - 1.750.000; (8) 1.750.001 - 2.000.000; (9) 2.000.001 - 2.250.000; (10) 2.250.001 - 2.500.000; (11) 2.500.001 - 2.750.000; (12) 2.750.001 - 3.000.000;
4. Atribut alamat pada Fig. 3 yang berisi data teks di rubah menjadi angka dengan menghitung jarak (JR) dari alamat mahasiswa menuju kampus dengan satuan kilometer (km).

Pada tahap pemodelan, memilih teknik pemodelan yang sesuai, alat penambangan data, algoritma yang digunakan, dan menyesuaikan aturan model untuk hasil yang optimal. Dalam penelitian ini menggunakan *K-Means* dan *K-Means++* sebagai algoritma dalam proses klasterisasi. Dalam perhitungan algoritma *K-Means* menggunakan fungsi *euclidian distance* yaitu perhitungan terhadap objek berdasarkan data terdekat sehingga dapat dihasilkan kelompok yang memiliki kemiripan dari setiap angotanya.

Proses evaluasi pada satu atau lebih model yang digunakan untuk menentukan nilai optimal untuk *k* menggunakan metode evaluasi DBI, *silhouette coefficient* dan metode *elbow*. Seluruh proses pemodelan dan evaluasi menggunakan aplikasi PyCharm dan *scikit-learn machine learning library* pada *Python*.

B. Data Mining

Data mining adalah proses menemukan pola dan trend yang berguna dalam kumpulan data yang besar [20], yang berkaitan dengan pengumpulan data, pemakaian data historis untuk menemukan pengetahuan, informasi, keteraturan, pola atau hubungan dalam data yang berukuran besar, output dalam data mining dapat dipergunakan sebagai alternatif dalam pengambilan keputusan atau untuk memperbaiki pengambilan keputusan di masa yang akan datang. Berdasarkan fungsi dan tujuan data mining di kelompokan menjadi deskripsi, klasifikasi, prediksi, estimasi, *clustering*, dan asosiasi[21].

Clustering merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidak miripan dengan *record* dalam kluster lain [22], yang mana kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal[23].

C. Algoritma K-Means

Algoritma K-Means mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien. Dengan kelebihan itu algoritma K-Means merupakan metode clustering yang umum di gunakan dalam berbagai penyelesaian masalah dalam kehidupan sehari hari [24].

Langkah-langkah pengelompokan data menggunakan fungsi *euclidean distance* dalam *K-Means* adalah sebagai berikut [6] :

1. Pilih jumlah *cluster* (k).
2. Inisialisasi awal pusat *cluster* dilakukan secara acak.
3. Setiap data ditempatkan ke pusat *cluster* terdekat sesuai jarak antar objek. Jarak dihitung berdasarkan kemiripan atau ketidak miripan data menggunakan metode jarak *euclidean* menggunakan rumus berikut:

$$d(x,y) = \sqrt{\sum_{i=n}^{n}(x_1 - y_1)^2} \quad (1)$$

dimana:
d (x,y) = ukuran ketidakmiripan
x_i = $(x_1, x_2, \ldots . x_n)$ variable data
y_i = $(y_1, y_2, \ldots . y_n)$ variable pada titik

4. Mengecek setiap data berdasarkan kedekatannya dengan jarak terkecil.
5. hitung pusat cluster yang baru dengan keanggotaan yang baru dengan cara menghitung rata-rata objek pada cluster.
6. Hitung kembali jarak tiap objek dengan pusat cluster yang baru, sehingga cluster tidak berubah, maka proses clustering selesai.

D. K-Means++

Algoritma K-means++ memiliki dua fase dimana pada fase pertama; *k-centroid* diidentifikasi tergantung pada nilai k yang telah dipilih secara umum ukuran jarak dihitung menggunakan jarak *Euclidian*. Sesuai Fig.3, pemilihan titik centroid awal ini diambil berdasarkan jarak terjauh antar titik centroid. Hal tersebut berdampak pada pengurangan jumlah perulangan pengelompokan data sehingga proses klasterisasi menjadi lebih cepat dan mudah. Fase kedua melibatkan penentuan centroid baru berdasarkan nilai rata-rata objek kelompok pada *cluster*. Proses perulangan dalam menemukan centroid baru sampai konvergensi terpenuhi.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i3.1439, Copyright ©2022
Submitted : July 19, 2022; Revised : January 7, 2023; Accepted : January 9, 2023; Published : January 25, 2023

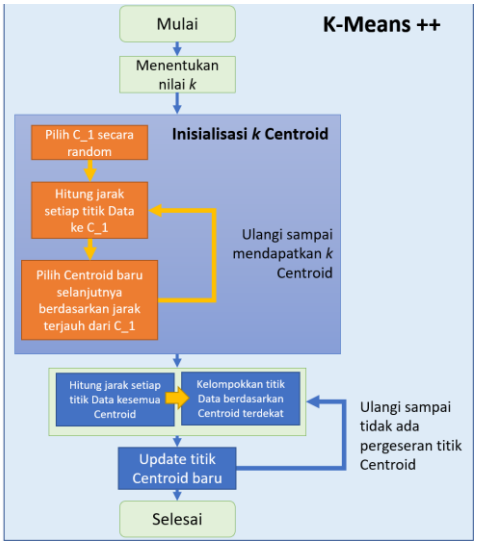

Fig. 3. Tahapan K-Means++

E. *Davies-Bouldin Index (DBI)*

DBI merupakan salah satu metode evaluasi internal yang mengukur evaluasi *cluster* pada suatu metode pengelompokan yang didasarkan pada nilai kohesi dan separasi. Dalam suatu pengelompokan, kohesi didefinisikan sebagai jumlah dari kedekatan data terhadap *centroid* dari *cluster* yang diikuti. Sedangkan separasi didasarkan pada jarak antar *centroid* dari *cluster*-nya[25]. Jika jarak intra-*cluster* minimal berarti masing-masing objek dalam *cluster* tersebut memiliki tingkat kesamaan karakteristik yang tinggi[26]. Semakin kecil nilai DBI yang diperoleh (non-negatif >= 0), maka semakin baik *cluster* yang diperoleh dari pengelompokan menggunakan algoritma *clustering*.

F. *Metode Elbow*

Klasterisasi K-Means meminimalkan jumlah kesalahan kuadrat (*sum squared error*/ sse) antara titik massa setiap klaster dan titik sampel dalam klaster sebagai tingkat distorsi. Pada sebuah klaster, jika distorsinya lebih rendah, koneksi antar anggota internalnya lebih dekat. Sebaliknya, semakin tinggi distorsi, semakin longgar struktur internalnya. Derajat distorsi akan berkurang jika jumlah cluster bertambah, dan kecepatan penurunan akan lebih lambat, ini titik tertentu sering dianggap sebagai titik dengan kinerja pengelompokan yang lebih baik[27]. Dan karena gambarnya menyerupai siku, maka dinamakan metode siku (*elbow*).

G. *Silhouette Coefficient*

Untuk validasi klaster, ada metrik internal yang disebut koefisien siluet yang memperhitungkan jarak intra-cluster dan antar-cluster. Kualitas pengelompokan dapat dinilai dengan menggunakan rata-rata siluet dimana memaksimalkan nilai indeks ini dapat digunakan untuk mencari jumlah cluster yang ideal[28].

III. HASIL DAN PEMBAHASAN

Pemodelan klasterisasi penelitian ini menggunakan algoritma *K-Means++* yang menghasilkan jumlah perulangan lebih sedikit dibanding *K-Means* versi inisial centroid acak. Terlihat pada Fig.4, perbandingan jumlah perulangan untuk ujicoba klasterisasi *K-Means* dan *K-Means++* dengan nilai k=1, 2, ... , 9, 10. Jumlah perulangan yang lebih sedikit berimplikasi pada waktu proses klasterisasi yang lebih cepat.

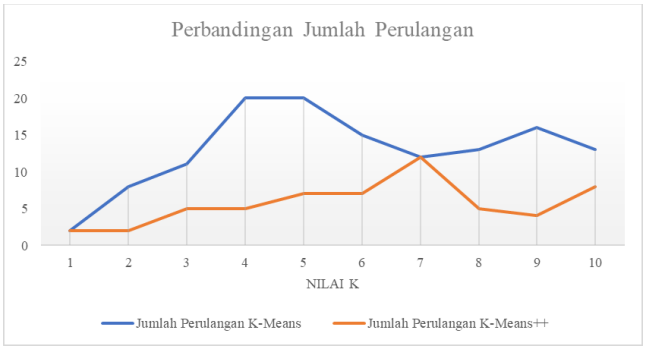

Fig. 4. Jumlah perulangan menggunakan K-Means dan K-Means++

Selanjutnya dilakukan pencarian model klasterisasi yang paling ideal berdasarkan nilai DBI, dengan mencoba nilai *k*=2, 3, ..., 9, 10. Pada Fig.5 dapat dilihat bahwa nilai DBI terkecil berada pada *k=2* dan *k=4*. Namun, jika melihat isi tabel II nilai DBI k=2 adalah 0.4905 sedangkan k=4 adalah 0.4929, sehingga dari nilai DBI yang terkecil adalah k=2.

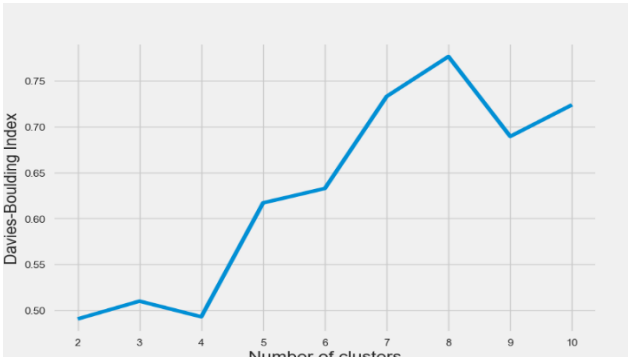

Fig. 5. Perbandingan Nilai DBI dengan K-Means++

Kemudian membandingkan nilai distorsi dari *sum squared error* (SSE) untuk pengelompokan data menggunakan nilai *k*= 1, 2, ..., 9, dan 10. *Sum of Square Error* (SSE) didapat dari mengukur selisih antara data yang diperoleh dengan model prediksi *K-Means++* untuk setiap nilai *k*[29]. Nilai-nilai tersebut digabungkan dan ditampilkan dalam bentuk grafik seperti pada Fig. 6, dimana terdapat bentuk siku (*elbow*) pada nilai k=2 dan k=3.

TABLE II. NILAI EVALUASI SILHOUETTE COEFFICIENT DAN DBI

k	2	3	4	5	6	7	8	9	10
Silhouette Coefficient	0.7523	0.6729	0.6645	0.5390	0.4806	0.3996	0.4022	0.4105	0.3941
Davies-Boulding Index	0.4905	0.5099	0.4929	0.6171	0.6328	0.7331	0.7767	0.6895	0.7240

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i3.1439, Copyright ©2022
Submitted : July 19, 2022; Revised : January 7, 2023; Accepted : January 9, 2023; Published : January 25, 2023

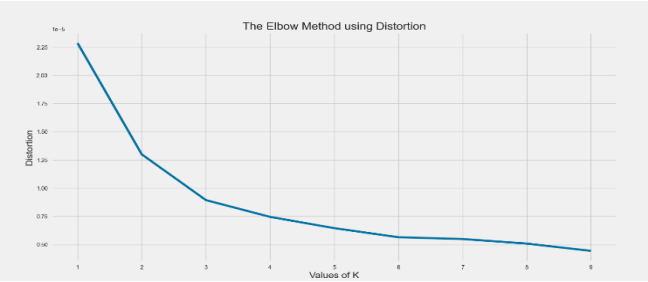

Fig. 6. Metode *Elbow* dengan nilai distorsi

Selanjutnya, evaluasi ketiga menggunakan *silhouette coefficient* yang ditampilkan pada Fig.7. Pada grafik tersebut ditemukan bahwa nilai terbesar terdapat pada k=2, dimana lebih rinci dari nilai *silhouette coefficient* tersebut dapat dilihat pada tabel II.

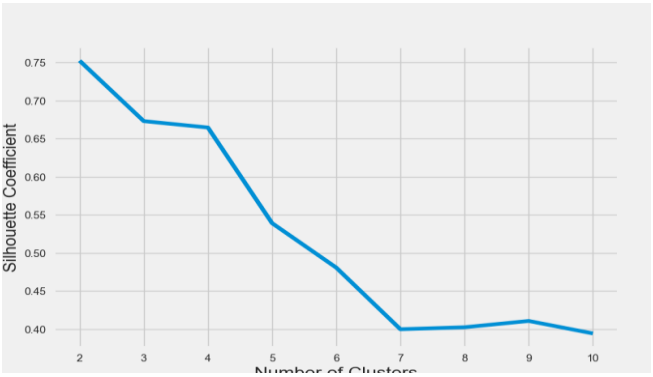

Fig. 7. Perbandingan Silhouette Coefficient dengan K-Means++

Berdasarkan hasil ujicoba diatas, dapat diketahui bahwa penentuan titik centroid diawal proses pengelompokan terbukti menurunkan jumlah perulangan pengelompokan data sesuai gambaran pada Fig. 4. Penggunanaan algoritma *K-Means* dengan inisial titik centroid yang acak membutuhkan perulangan yang lebih banyak khususnya pada k=4 dan k=5 yang mencapai 20 kali perulangan. Namun, penggunaan *K-Means++* yang menentukan titik centroid berdasarkan jarak terjauh dengan titik centroid pertama, hanya memerlukan jumlah perulangan yang lebih sedikit. Walau pada satu kondisi penggunaan kedua algoritma ini memerlukan jumlah perulangan yang sama, yaitu saat k=7. Hal ini, memungkinkan terjadi karena pada *K-Means++* penentuan titik centroid pertama masih ditentukan secara acak. Namun, penggunaan *K-Means++* pada dataset yang berskala besar dapat membantu proses klasterisasi menjadi lebih cepat karena jumlah perulangannya lebih sedikit dibandingkan penggunaan *K-Means* dengan inisialisasi titik centroid yang random[30].

Selanjutnya melihat dari Fig. 6, dari metode *elbow* dalam menentukan nilai *k* optimal harus dapat mengidentifikasi titik siku berada di nilai *k* berapa, namun pada gambar *k*=2 dan *k*=3 memungkinkan untuk dipilih menjadi nilai optimal. Namun dengan memperhatikan nilai dari *silhouette coefficient* dan DBI, klasterisasi dengan *k*=2 dan k=3 masing-masing memiliki nilai *silhouette coefficient* adalah 0.7523 dan 0.6729 (lihat Tabel II), dan nilai DBI masing-masing adalah 0.4905 dan 0.5099 (lihat Tabel II). Nilai *silhouette coefficient* untuk model *k*=2 lebih besar dibanding model *k*=3, dan nilai DBI *k*=2 lebih kecil dari model yang lainnya. Sehingga berdasarkan evaluasi nilai SSE (metode *elbow*), *silhouette coefficient* dan DBI, model yang memiliki pengelompokan paling optimal adalah model klasterisasi dengan nilai *k*=2.

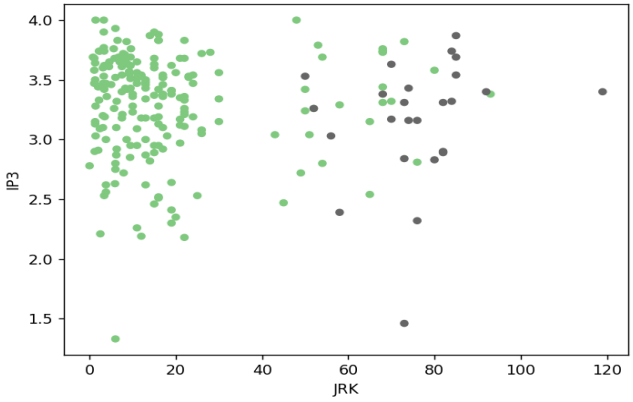

Fig. 8. Hasil Klasterisasi K-Means++ dengan *k*=2

Pada Fig.8 dapat dilihat bahwa terdapat dua area sebaran data, dimana titik data berwarna hijau menandai anggota klaster ke-1 dan titik data berwarna hitam menandai anggota klaster ke-2. Dengan menampilkan pada grafik dua dimensi, dipilih atribut non-akademik yaitu jarak (JRK) dan atribut akademik yaitu IP semester 3 (IP3). Atribut jarak merupakan atribut yang ditambahkan pada penelitian ini, yang mana pada penelitian mengenai pemetaan karakteristik mahasiswa belum ada yang menggunakannya[5][6][7][8][9][10]. Berdasarkan penelitian ini, maka atribut jarak tempat tinggal mahasiswa ke lokasi kampus layak untuk dijadikan sebagai atribut non-akademik dalam pemetaan karakteristik mahasiswa, selaras dengan penelitian sebelumnya bahwa karakteristik mahasiswa non-akademik salah satunya lingkungan tempat tinggal[2][3].

Selanjutnya hasil dari klasterisasi pada dataset yang digunakan, didapatkan pemetaan karakteristik mahasiswa terhadap 2 klaster yang memiliki nilai rata-rata setiap atribut seperti yang ditampilkan pada Tabel III.

TABLE III. NILAI EVALUASI SILHOUETTE COEFFICIENT DAN DBI

	JK	PK	JT	PH	PR	JRK	IP1	IP2	IP3	IP4
K_1	1.48	4.78	4.89	1,408,359	0.15	16.21	3.35	3.31	3.29	3.31
K_2	1.33	5.00	4.93	882,000	0.13	74.17	3.30	3.16	3.22	3.19

Pada Tabel II diatas terdapat perbedaan nilai IP dimana *cluster* satu memiliki nilai IP yang lebih besar dibandingkan dengan nilai IP *cluster* dua. Perbandingan nilai yang paling signifikan yaitu pada atribut jarak rumah menuju kampus (JR). Rata-rata jarak rumah ke kampus pada *cluster* satu lebih kecil dibandingkan nilai pada cluster dua. Melihat kondisi tersebut, maka dari pemetaan ini dapat menunjukan dua karakteristik mahasiswa yang berbeda baik dari faktor akademik dan non-akademik.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i3.1439, Copyright ©2022
Submitted : July 19, 2022; Revised : January 7, 2023; Accepted : January 9, 2023; Published : January 25, 2023

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai implementasi *K-Means* dalam pemetaan karakteristik mahasiswa penerima KIP-K ini, maka dapat diambil kesimpulan bahwa: 1) algoritma *K-Means++* memiliki jumlah perulangan yang lebih sedikit dibanding *K-Means*, sehingga proses menjadi lebih cepat; 2) model klasterisasi *K-Means++* dengan dataset yang dipilih, memiliki nilai *k* optimal yaitu *k*=2 berdasarkan evaluasi metode *elbow* dengan *silhouette coefficient* = 0.7523, dan DBI-nya = 0.49053; serta 3) model klasterisasi yang dibangun telah mampu memetakan karakteristik mahasiswa penerima KIP-K pada 2 kelompok dengan atribut yang signifikan berbeda yaitu IP semester 1 sampai semester 4 dan jarak rumah ke lokasi kampus.

Untuk pengembangan penelitian yang serupa dikemudian hari, disarankan untuk menambahkan atribut diperlukan untuk mengetahui lebih banyak faktor yang mempengaruhi nilai IP mahasiswa penerima KIP-K seperti faktor kegiatan non-akademik yang dilaksanakan di kampus dan luar kampus.

REFERENCES

[1] M. M. Manurung and R. Rahmadi, "Identifikasi Faktor-faktor Pembentukan Karakter Mahasiswa," *JAS-PT J. Anal. Sist. Pendidik. Tinggi*, vol. 1, no. 1, p. 41, 2017, doi: 10.36339/jaspt.v1i1.63.

[2] R. Rosmini, A. Fadlil, and S. Sunardi, "Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah," *It J. Res. Dev.*, vol. 3, no. 1, pp. 22–31, 2018, doi: 10.25299/itjrd.2018.vol3(1).1773.

[3] A. N. Syaharani and F. Nurani, "Kesenjangan Mutu Pendidikan Antara Desa dan Kota," 2019.

[4] I. Pratama and P. T. Prasetyaningrum, "Pemetaan Profil Mahasiswa Untuk Peningkatan Strategi Promosi Perguruan Tinggi Menggunakan Predictive Apriori," *J. Eksplora Inform.*, vol. 10, no. 2, pp. 159–166, 2021, doi: 10.30864/eksplora.v10i2.505.

[5] D. Kurniadi, E. Abdurachman, H. L. H. S. Warnars, and W. Suparta, "The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 434, no. 1, 2018, doi: 10.1088/1757-899X/434/1/012039.

[6] A. E. Rahayu, K. Hikmah, N. Y. Ningsih, and A. C. Fauzan, "Penerapan K-Means Clustering Untuk Penentuan Klasterisasi Beasiswa Bidikmisi Mahasiswa," *Comput. Sci. Appl. Informatics*, vol. 1, no. 2, pp. 82–86, 2019.

[7] E. Buulolo, R. Syahputra, and A. Fau, "Algoritma K-Medoids Untuk Menentukan Calon Mahasiswa Yang Layak Mendapatkan Beasiswa Bidikmisi di Universitas Budi Darma," vol. 4, pp. 797–805, 2020, doi: 10.30865/mib.v4i3.2240.

[8] A. Hardianti and D. Agushinta R, "ANALISIS POLA MASA STUDI MAHASISWA FAKULTAS TEKNIK UNIVERSITAS DARMA PERSADA MENGGUNAKAN METODE CLUSTERING K-MEANS," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 4, pp. 861–868, 2020, doi: 10.25126/jtiik.202071001.

[9] S. Astuti, Samsudin, and Triase, "Penerapan Data Mining Dalam Menentukan Penerima Beasiswa Upz (Unit Pengumpulan Zakat) Menggunakan Algoritma K-Means," vol. 13, no. 2, 2021.

[10] I. Irmayansyah and S. E. Triyono, "Penerapan Algoritma K-Means Untuk Pemetaan Potensi Calon Mahasiswa Baru," *Teknois J. Ilm. Teknol. Inf. dan Sains*, vol. 12, no. 2, pp. 139–150, 2022, doi: 10.36350/jbs.v12i2.139.

[11] H. Gunawan and V. Purwayoga, "DATA MINING MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING UNTUK MENGETAHUI POTENSI PENYEBARAN VIRUS CORONA DI KOTA CIREBON," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 1, pp. 1–8, Jan. 2022, doi: 10.32736/sisfokom.v11i1.1316.

[12] A. Darussalam, "Perbandingan Akurasi Metode Clustering Algoritma K-Means Dengan Algoritma K-Medoids Dalam Pengelompokan Data Mahasiswa Baru Untuk Strategi Promosi Program Studi Teknik Informatika Unisnu Jepara," vol. 3, pp. 1–9, 2019.

[13] F. N. R. F. Aziz, B. D. Setiawan, and I. Arwani, "Implementasi Algoritma K-Means untuk Klasterisasi Kinerja Akademik Mahasiswa," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 6, pp. 2243–2251, 2018.

[14] F. Nuraeni and L. Listiani, "Implementation of K-Means Algorithm with Distance of Euclidean Proximity in Clustering Cases of Violence Against Women and Children," in *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, 2019, no. August, pp. 162–167.

[15] S. Surohman, L. Fabrianto, F. Riza, and N. M. Faizah, "Korelasi Antara Profil dan Nilai Akademis Siswa dengan Menggunakan Algoritma K-Means," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 4, p. 845, 2021, doi: 10.25126/jtiik.2021843034.

[16] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, vol. 56, no. 3, pp. 1283–1291, Jun. 2022, doi: 10.1007/s11135-021-01176-w.

[17] F. Nuraeni, D. Tresnawati, Y. H. Agustin, and G. F. Dermawan, "OPTIMIZATION OF MARKET BASKET ANALYSIS USING CENTROID-BASED CLUSTERING ALGORITHM AND FP-GROWTH ALGORITHM OPTIMALISASI ANALISIS KERANJANG PASAR MENGGUNAKAN ALGORITMA CENTROID-BASED CLUSTERING DAN ALGORITMA FP-GROWTH," *J. Tek. Inform.*, vol. 3, no. 6, pp. 1581–1590, 2022, doi: https://doi.org/10.20884/1.jutif.2022.3.6.399 p-ISSN:

[18] C. Ayudia, S. Fastaf, and Y. Yamasari, "Analisa Pemetaan Kriminalitas Kabupaten Bangkalan Menggunakan Metode K-Means dan K-Means ++," *J. Informatics Comput. Sci.*, vol. 03, no. 04, pp. 534–546, 2022.

[19] A. Kapoor and A. Singhal, "A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms," in *3rd IEEE International Conference on*, 2017, pp. 1–6, doi: 10.1109/CIACT.2017.7977272.

[20] D. T. Larose and C. D. Larose, *Discovering Knowledge In Data An Introduction To Data Mining Second Edition Wiley Series On Methods And Applications In Data Mining*. 2014.

[21] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*. Yogyakarta: DEEPUBLISHER, 2020.

[22] M. Wahyudi, Mashita, R. Saragih, and Solikhun, *Data Mining Penerapan Algoritma K-Means Clustering dengan K-Medoids Clustering*. Yayasan Kita Menulis, 2020.

[23] 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, doi: 10.25077/teknosi.v3i1.2017.211-218.

[24] Y. Amri, "Metode k-means untuk clustering mahasiswa berdasarkan nilai akademik," vol. XV, no. 02, 2021.

[25] Z. Nabila, A. Rahman Isnain, and Z. Abidin, "Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 2, p. 100, 2021.

[26] B. Jumadi and USU, "Peningkatan Hasil Evaluasi Clustering Davies Bouldin dengan Penentuan Titik Pusat Cluster Awal K-Means," 2018.

[27] X. Guo, "Clustering of NASDAQ Stocks Based on Elbow Method and K-Means," in *Proceedings of the 4th International Conference on Economic Management and Green Development*, 2021, pp. 80–87, doi: 10.1007/978-981-16-5359-9_11.

[28] D.-T. Dinh, T. Fujinami, and V.-N. Huynh, "Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient," in *20th International Symposium, KSS 2019*, 2019, pp. 1–17, doi: 10.1007/978-981-15-1209-4_1.

[29] R. Nainggolan, R. Perangin-Angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *J. Phys.*

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i3.1439, Copyright ©2022
Submitted : July 19, 2022; Revised : January 7, 2023; Accepted : January 9, 2023; Published : January 25, 2023

Conf. Ser., vol. 1361, no. 1, 2019, doi: 10.1088/1742-6596/1361/1/012015.

[30] J. Hämäläinen, T. Kärkkäinen, and T. Rossi, "Improving scalable k-means++," *Algorithms*, vol. 14, no. 1, pp. 1–20, 2021, doi: 10.3390/a14010006.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v11i3.1439, Copyright ©2022
Submitted : July 19, 2022; Revised : January 7, 2023; Accepted : January 9, 2023; Published : January 25, 2023