Implementasi Algoritma K-Modes untuk Mendukung Pengambilan Keputusan Rencana Akademis Mahasiswa

Andrew Widjaya

School Of Computer Science

Binus University

Jakarta, Indonesia

andrew032@binus.ac.id

Kevina Nugraha Eleeas

School Of Computer Science

Binus University

Jakarta, Indonesia

kevina.eleeas@binus.ac.id

Gabrielle Felicia Ariyanto

School Of Computer Science

Binus University

Jakarta, Indonesia

gabrielle.ariyanto@binus.ac.id

Jocelyn Verna Siswanto

School Of Computer Science

Binus University

Jakarta, Indonesia

jocelyn.siswanto@binus.ac.id

Abstrak - Perencanaan akademik merupakan layanan yang diberikan oleh para penasihat akademik untuk membantu mahasiswa dalam mempersiapkan dan menyelesaikan perkuliahan secara sistematis, terukur, dan terencana hingga mampu lulus tepat waktu. Penasehat akademik sendiri bertugas untuk mendorong mahasiswa mencapai prestasi akademik demi mempersiapkan karir kedepan sesuai dengan kemampuan masing-masing[1]. Faktanya, banyak rencana akademik yang terkendala dan gagal karena berbagai faktor, salah satunya perbedaan karakteristik mahasiswa. Penelitian ini bertujuan untuk membantu penasihat akademik dalam menentukan kelompok-kelompok mahasiswa yang membutuhkan bimbingan akademik yang lebih intens agar mahasiswa dapat mencapai sasaran yaitu lulus tepat waktu berdasarkan analisa karakteristik mahasiswa menggunakan metode algoritma *K-Modes*. Algoritma *K-Modes* adalah algoritma *clustering* yang cocok digunakan pada data yang bersifat kategorikal [2]. Dalam menentukan jumlah *cluster* yang terbaik, kami menggunakan metode *evaluation metrics* berupa *Elbow Method* dan juga *Calinski-Harabasz Index (CH Index)*. Jumlah *cluster* yang terbaik adalah jumlah *cluster* yang memiliki *Index CH value* maksimum, dimana pada makalah ini terdapat 4 *cluster*, yakni *cluster* 0 sejumlah 3.931 data, *cluster* 1 sejumlah 3.141 data, *cluster* 2 sejumlah 2.251 data, serta *cluster* 3 sejumlah 1.951 data.

Kata Kunci - Rencana Akademik, Mahasiswa, Cluster, K-Modes Clustering, Calinski-Harabasz Index

I. Pendahuluan

i. Latar Belakang

Seorang mahasiswa dapat menyelesaikan masa studi mereka dengan waktu yang bervariasi, mulai dari 3,5 tahun (lulus lebih cepat), 4 tahun (lulus tepat waktu), serta lebih dari 4 tahun (lulus terlambat). Sayangnya, di banyak institusi publik luar negeri, lebih banyak mahasiswa yang terlambat menyelesaikan masa studinya (40% menyelesaikan masa studinya dalam kurun waktu 3,5 - 4 tahun, sedangkan hampir 60% sisanya menyelesaikan masa studi dalam kurun waktu 6 tahun)[3]. Hal ini sungguh disayangkan, sebab dengan tidak lulus tepat waktu, banyak uang dan tenaga yang terbuang sia-sia untuk menyelesaikan masa studi tersebut.

Dosen pembimbing merupakan salah satu elemen penting dalam membantu mahasiswa dapat menjalani masa studi mereka secara mulus selama periode pembelajarannya[4]. Maka dari itu, kontribusi dosen diperlukan dalam membuat sebuah rencana akademis yang efektif dan tepat sasaran guna membantu mahasiswa lulus tepat waktu. Dosen Pembimbing Akademik dan Ketua Program Studi akan mengambil langkah dan juga inisiatif yang tepat dalam menyusun rencana studi mahasiswa agar dapat lulus tepat waktu.

Mahasiswa yang masa studinya terlambat pun dapat dimaksimalkan agar dapat lulus secepat mungkin.

Namun, tidak semua rencana akademik berjalan dengan baik. Banyak rencana akademik yang berujung gagal dan terkendala yang disebabkan oleh beberapa faktor, terutama karakteristik mahasiswa. Karakteristik mahasiswa, seperti keaktifan berorganisasi, jenis UKM yang diikuti, status bekerja, dan lain sebagainya cukup beragam di kalangan mahasiswa sehingga diperlukan pengelompokkan mahasiswa berdasarkan karakteristiknya.

Pada makalah ini, akan dibuat pengelompokkan mahasiswa berdasarkan karakteristiknya dengan menggunakan algoritma *clustering K-Modes*, dengan validasi jumlah *cluster* yang optimal menggunakan metode *Elbow* dan *Calinski-Harabasz* sebagai metrik evaluasi yang digunakan. Dengan ini, diharapkan agar makalah ini dapat menjadi sebuah *supporting decision making* bagi penasihat akademik agar dapat menentukan kelompok mahasiswa yang perlu untuk diberi perhatian, pelatihan, dorongan, dan bimbingan yang lebih intens guna mengoptimalkan masa studi mahasiswa dengan rencana akademik yang tepat.

ii. Rumusan Masalah

Berdasarkan latar belakang masalah maka dapat dirumuskan suatu masalah dijabarkan sebagai berikut:

1. Bagaimana karakteristik tiap-tiap kelompok mahasiswa?
2. Bagaimana karakteristik kelompok mahasiswa yang membutuhkan bimbingan yang lebih intens karena terkendala, bahkan terlambat lulus?

iii. Tujuan dan Manfaat

Adapun tujuan dan manfaat penelitian ini adalah sebagai berikut:

1. Memaparkan dan mendeskripsikan karakteristik mahasiswa
2. Mendiagnosis apa saja faktor penghambat mahasiswa dalam menjalani dan menyelesaikan perkuliahan agar dapat diberikan bimbingan yang lebih intens

II. Pembahasan

i. Landasan Teori

Menurut Kamus Besar Bahasa Indonesia (KBBI), mahasiswa didefinisikan sebagai orang yang belajar di Perguruan Tinggi [5]. Berdasarkan peraturan pemerintah RI No. 30 tahun 1990, mahasiswa adalah peserta didik yang terdaftar dan belajar di perguruan tinggi tertentu [6]. Masa studi mahasiswa S1 berkisar dari 4 sampai 7 tahun, dimana mahasiswa dapat dikatakan lulus tepat waktu jika menyelesaikan masa studi dalam 4 tahun atau 8 semester. Namun, mahasiswa juga dapat mempercepat masa studinya menjadi 3,5 tahun saja atau semester jika memenuhi syarat untuk menjalankannya.

Keberhasilan mahasiswa untuk lulus juga perlu diperhatikan oleh setiap institusi perguruan tinggi. Adanya dosen pembimbing akademik yang menyediakan konsultasi bimbingan akademik merupakan salah satu upaya untuk meningkatkan peluang mahasiswa lulus tepat waktu. Untuk mendapatkan hasil yang maksimal, dosen pembimbing perlu mengetahui faktor faktor yang menunjukkan karakteristik mahasiswa yang dibimbingnya, sebab setiap mahasiswa memiliki cara dan solusi yang berbeda dalam menghadapi masa studinya.

Pendekatan *Machine Learning* dapat digunakan untuk memecahkan pengelompokkan mahasiswa berdasarkan karakteristiknya. *Unsupervised learning* merupakan salah satu algoritma *Machine Learning* yang digunakan untuk mempelajari data berdasarkan kedekatannya dengan data yang lain. Salah satu implementasi dari *Unsupervised learning* ini adalah *Clustering*. *Clustering* adalah teknik partisi dengan mengelompokan sekumpulan objek menjadi beberapa kelompok kelompok yang disebut *cluster* dimana objek di dalam satu *cluster* mirip satu sama lain tetapi berbeda dengan objek di *cluster* lain . *Clustering* merupakan metode segmentasi data yang berfungsi untuk memprediksi dan menganalisa suatu masalah bisnis [7].

Algoritma *clustering k-Modes* merupakan pengembangan dari algoritma *k-Means* untuk pengelompokan data kategorik yang pertama kali diperkenalkan oleh Huang pada tahun 1997. Algoritma *clustering k-Means* standar tidak dapat diaplikasikan untuk data kategorik karena fungsi jarak *euclidean* dan penggunaan rata-rata untuk merepresentasikan pusat *cluster* .

Metode *Elbow Method* merupakan salah satu metode yang paling umum untuk mencari jumlah *cluster* yang optimum [8]. Metode *Elbow* melakukan perhitungan dan perbandingan nilai *sum of square error* (SSE) pada setiap *cluster*. Semakin banyak jumlah *cluster* maka semakin rendah nilai SSE yang akan kita peroleh.

Rumus dari *sum of square error* (SSE) sendiri adalah sebagai berikut:

$$\text{SSE} = \sum_{k=1}^{k} \sum_{x_i \in S_k} \left\| X_i - C_k \right\|^2$$

Keterangan:

K : Jumlah *cluster* yang digunakan pada algoritma K-Means.

X_i : Nilai atribut data ke-i

C_k : Nilai atribut titik pusat *cluster* ke-k.

Menentukan jumlah *cluster* yang optimum juga dapat dilakukan dengan melihat grafik metode *elbow.* Grafik dibuat dengan menentukan jumlah SSE dengan jumlah *cluster* dari range 1 sampai 10. Hasil *cluster* yang optimum terdapat pada titik dimana berbentuk siku, yang berarti terdapat penurunan nilai SSE yang drastis.

Calinski-Harabasz index yang dikenal juga sebagai '*the Variance Ratio Criterion*' adalah metode penghitungan sebagai rasio jumlah dispersi *antar-cluster* dan jumlah dispersi *intra-cluster* untuk semua *cluster*, dimana dispersi adalah jumlah kuadrat jarak. *Calinski-Harabasz* digunakan untuk menentukan jumlah cluster yang optimal dimana *Calinski-Harabasz Index* yang tinggi menunjukkan pengelompokan yang lebih baik karena pengamatan objek di setiap *cluster* lebih dekat satu sama lain, namun jauh dengan cluster yang lain. [9]. *Calinski Harabasz Index* dapat dihitung dengan perhitungan berikut

$$CH = SSB/SSW * (N - K)/(K - 1)$$

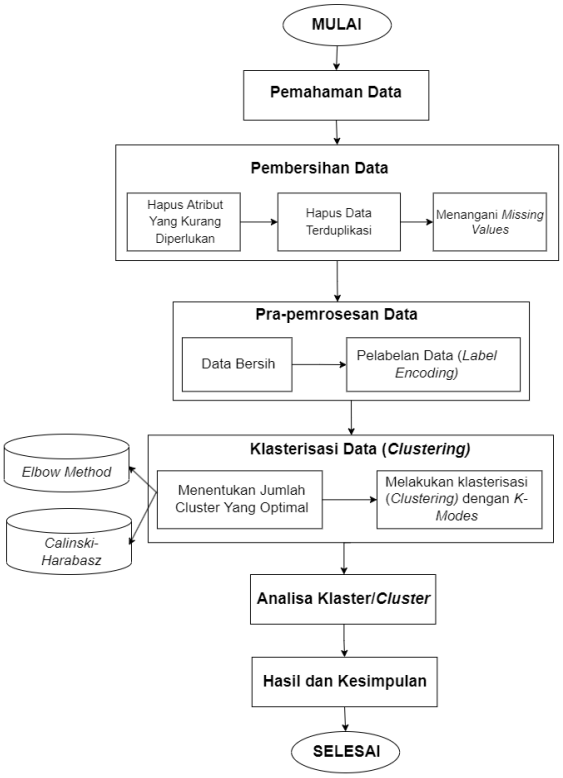
Keterangan :

- SSB = *Sum of Squared Between*
- SSW = *Sum of Squared Within*
- N = Jumlah Data
- K = Jumlah *cluster* yang digunakan pada algoritma

ii. Metode Penelitian

Dataset yang digunakan dalam penelitian ini adalah data yang telah diberikan oleh penyelenggara *IT Competition Informatics Festival (IFEST)* 2022 dalam ajang perlombaan *Data Analytics Competitions* tentang mahasiswa di satu universitas yang sama dengan karakteristiknya yang dapat diakses melalui link yang tertera di bawah bit.ly/DatasetBabakPenyisihanDAC2022.

Dalam penelitian ini, kami menggunakan bahasa pemrograman Python dengan *Jupyter Notebook* sebagai software untuk *Cleaning, Pre-Processing*, *Clustering* dan analisis data. Alur kerja dapat dilihat dalam *Flowchart* yang terdapat pada Gambar-1.



Gambar-1 : Sistem Kerja

A. Pemahaman Data

Pemahaman Data atau *Data Understanding* merupakan tahapan penting dalam *Data Mining*, dimana *Data Understanding* membantu kita untuk mengetahui karakteristik awal dari dataset, sehingga kita lebih familiar terhadap data yang ingin dianalisis[10]. Pada tahap pemahaman data, kami melakukan analisa data untuk melihat jumlah baris dan kolom, serta nama kolom yang terdapat dalam dataset.

B. Pembersihan Data

Pembersihan data atau *Data Cleaning* adalah suatu proses untuk mengubah data mentah yang inkonsisten, tidak bersih (memiliki *missing values)* menjadi data yang konsisten agar dapat dianalisis lebih lanjut[11]. Adapun pembersihan data yang kami lakukan adalah:

- Hapus Atribut Yang Kurang Diperlukan, yaitu menghapus atribut atau variabel yang kurang berguna untuk mencapai tujuan penelitian kami. Pada tahap ini, diputuskan bahwa atribut 'Nama' tidak memiliki pengaruh dalam *clustering,* sehingga atribut 'Nama' dihapus dalam dataset.

- Hapus Data Terduplikasi, yaitu menghapus baris data yang berulang.
- Menangani *missing Value.* Adanya *missing value* yang tidak ditangani dapat mengganggu proses analisa pada data seperti adanya data yang tidak terdistribusi secara normal, adanya prefiks, affiks, ataupun suffiks.

  Tahapan yang dilakukan adalah mencari *missing value* pada dataset dan didapati variabel yang mengandung *missing value* adalah variabel 'Biaya' sebanyak 3.779 data dan 'Tinggal_Dengan' sebanyak 3.982 data. Setelah diamati, kami melakukan re-inputasi kedua variabel tersebut dengan sebuah kategori baru yaitu 'Unknown', sebab kedua variabel tersebut memiliki informasi yang cukup menarik untuk digali, serta persentase *missing value* yang cukup besar , yakni >20% dari keseluruhan data, membuat kami memutuskan untuk tidak langsung menghapus kolom yang mengandung *missing value* tersebut.

C. Pra-pemprosesan Data

Pada tahap pra-pemrosesan data (*Data Preprocessing*) dilakukan proses untuk mengubah data yang sudah dibersihkan pada tahap sebelumnya menjadi data yang lebih bermanfaat dan informatif.

- Data bersih → Pelabelan Data (*Label Encoding*)

Di tahap pra-pemrosesan data, kami melakukan pelabelan data (*Label Encoding*) pada data yang sudah dibersihkan. *Label encoding* adalah teknik untuk membantu proses *Clustering* data kategorikal pada tahap berikutnya dengan cara mengubah kategori menjadi angka[12].

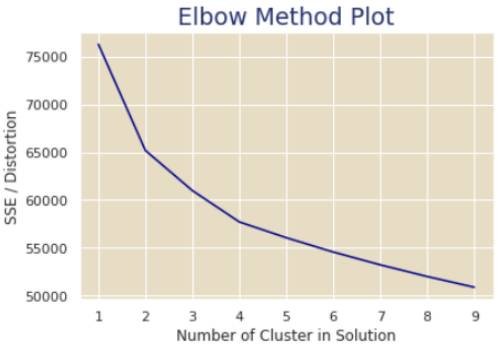
D. Klasterisasi Data (*Clustering*)

Clustering Data merupakan analisis pengelompokkan dengan memasukkan titik data ke dalam kelompok tertentu agar dapat memperoleh wawasan berharga dari data kita[13]. Adapun metode *clustering* yang kami gunakan adalah Algoritma *clustering K-Modes*. Langkah-langkah dalam algoritma *clustering k-Modes* adalah:

1. Memilih data inisiasi awal sebagai titik pusat, satu untuk setiap *cluster* dengan metode 'Cao' yang menggunakan algoritma *MaxMin* dengan mengembangkan penggunaan rata-rata frekuensi untuk menggantikan inisiasi secara acak [14] .
2. Menghitung jarak masing-masing data terhadap semua titik pusat *cluster*. Alokasikan setiap objek ke *cluster* terdekat menggunakan ukuran ketidaksamaan sederhana.
3. Setelah semua objek dialokasikan ke *cluster*, lakukan pengujian ulang perbedaan objek terhadap modus. Jika objek lebih mendekati *cluster* lain daripada *cluster* saat ini, maka alokasikan ulang objek ke *cluster* tersebut dan perbaharui modus kedua *cluster*.
4. Mengulangi langkah ketiga hingga tidak ada objek yang berubah *cluster* setelah dilakukan satu iterasi penuh terhadap seluruh data

Selanjutnya, kami menggunakan metode *Elbow* (*Elbow Method*) sebagai *evaluation metrics* yang bertujuan untuk menentukan jumlah cluster yang optimum atau yang terbaik. Adapun, langkah-langkah yang dilakukan pada metode *Elbow* adalah :

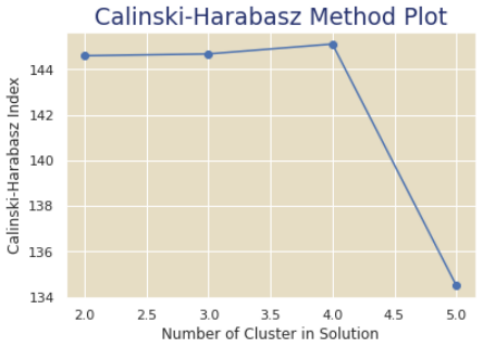
1. Membuat variasi nilai *cluster*, yaitu 1 sampai 10.
2. Menghitung nilai *sum of squares error* (SSE) dari tiap-tiap *cluster* sampai jumlah *cluster* yang ditentukan.
3. Membuat grafik perhitungan *sum of squares error* (SSE) yang sudah dilakukan sebelumnya.
4. Membandingkan nilai *sum of squares error* (SSE) dan mengambil nilai *cluster* yang turun secara drastis/signifikan. Dengan kata lain, mengambil nilai yang berbentuk siku pada grafik.



Gambar-2 : Grafik *Elbow Method Plot*

Berdasarkan grafik diatas (Gambar-2), dapat dinyatakan bahwa jumlah *cluster* yang optimum menggunakan metode *Elbow* adalah 4 *cluster* ,dimana terjadi penurunan SSE yang tajam pada number of *cluster* 4 yang menunjukkan "siku" dari grafik tersebut .

Kemudian untuk *evaluation metrics* kedua, kami menggunakan *Calinski - Harabasz Score* (CH *Score*) dengan membandingkan CH Skor dari tiap tiap variasi jumlah *cluster*(k). CH Skor tertinggi menandakan bahwa *cluster* memiliki nilai kesamaan yang tinggi.



Gambar-3 : Grafik *Calinski-Harabasz Method Plot*

Grafik diatas menunjukkan pemetaan metode *Calinski-Harabasz* dengan tiap-tiap jumlah *cluster* untuk menentukan jumlah *cluster* yang optimal. Berdasarkan grafik *Calinski-Harabasz* (dapat dilihat di Gambar-3), tampak index tertinggi yang ditunjukan di sumbu Y memiliki 4 *cluster* yang ditunjukan di sumbu X.

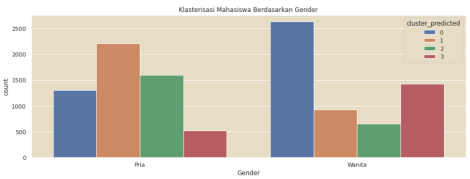
iii. Hasil dan Pembahasan

1. Penentuan *cluster* optimal dan *Clustering*

   Berdasarkan hasil grafik *Elbow Method Plot* (Gambar-2), jumlah *cluster* yang optimum pada data mahasiswa ini adalah 4. Hal ini diperkuat juga dengan metode *Calinski-Harabasz* yang kami gunakan, dimana pada hasil grafik *Calinski-Harabasz* (Gambar-3) index tertingginya ditunjukkan pada angka 4 yang berarti jumlah *cluster* yang terbaik adalah 4 *cluster*. Kemudian, dengan menggunakan bahasa pemrograman Python, kami melakukan *Clustering* dengan inisiasi awal menggunakan metode 'Cao' dan menghasilkan 4 *cluster* yang kemudian kami visualisasi dan analisa per *cluster*.

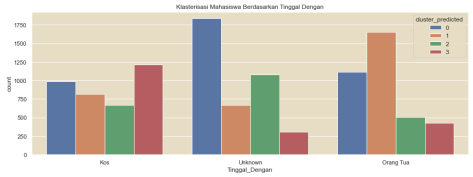
2. Visualisasi Hasil *Clustering*

   a. Berdasarkan Gender



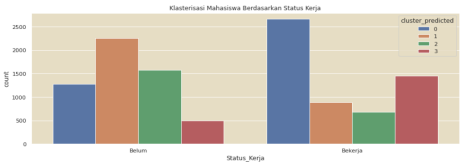
Gambar-4: Plot *Clustering* Mahasiswa Berdasarkan Gender

b. Berdasarkan Tinggal Dengan



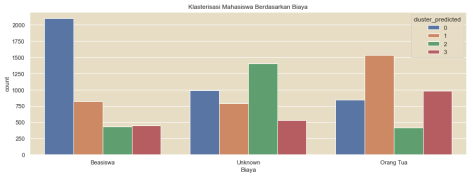
Gambar-5: Plot *Clustering* Mahasiswa Berdasarkan Tinggal Dengan

c. Berdasarkan Status Kerja



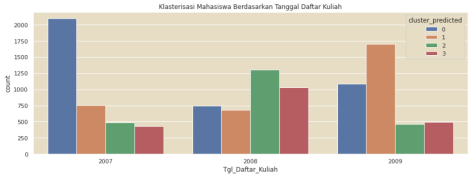
Gambar-6: Plot *Clustering* Mahasiswa Berdasarkan Status Kerja

d. Berdasarkan Biaya



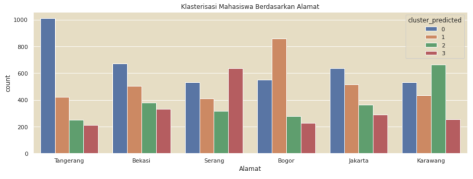
Gambar-7: Plot *Clustering* Mahasiswa Berdasarkan Biaya

e. Berdasarkan Tanggal Daftar Kuliah



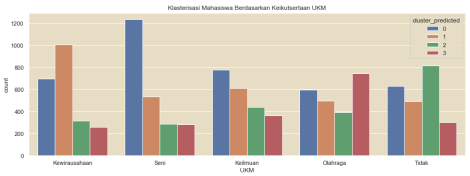
Gambar-8: Plot *Clustering* Mahasiswa Berdasarkan Tanggal Daftar Kuliah

f. Berdasarkan Alamat



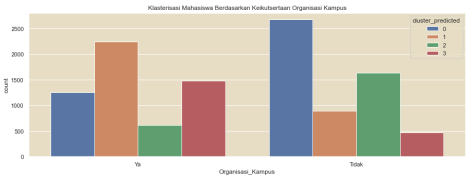
Gambar-9: Plot *Clustering* Mahasiswa Berdasarkan Alamat

g. Berdasarkan Keikutsertaan Unit Kegiatan Mahasiswa



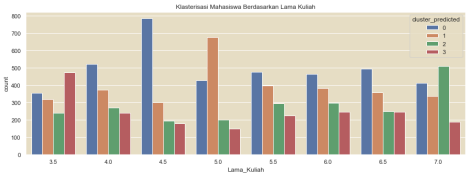
Gambar-10: Plot *Clustering* Mahasiswa Berdasarkan Keikutsertaan UKM

h. Berdasarkan Keikutsertaan Organisasi Kampus



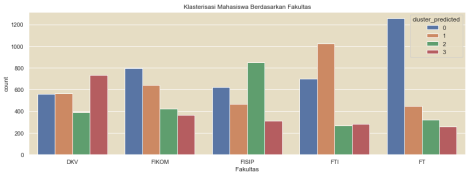
Gambar-11: Plot *Clustering* Mahasiswa Berdasarkan Keikutsertaan Organisasi Kampus

i. Berdasarkan Lama Kuliah



Gambar-12: Plot *Clustering* Mahasiswa Berdasarkan Lama Kuliah

j. Berdasarkan Fakultas



Gambar-13: Plot *Clustering* Mahasiswa Berdasarkan Fakultas

3. Analisa *Cluster*

Setelah membaca visualisasi *cluster* dan sentroid tiap *cluster*, berikut adalah tabel identifikasi dari tiap-tiap *cluster*

Cluster	Penjelasan
0	*Cluster* ini memiliki jumlah anggota terbanyak dari *cluster* yang lain. Dengan jumlah **3.931 anggota**, *cluster* ini **didominasi anggota wanita** daripada pria. Walaupun kebanyakan anggota tidak diketahui status tinggalnya, persebaran anggota ***cluster* yang tinggal kos dan dengan orangtua cukup merata**. Anggota *cluster* 0 didominasi oleh mahasiswa yang **mendaftar kuliah pada tahun 2007**. *Cluster* ini juga didominasi oleh mahasiswa yang **sudah bekerja dan tidak berorganisasi** pada saat berkuliah. Mayoritas mahasiswa di *cluster* 0 mengambil **fakultas FT** dan juga **menerima beasiswa** dengan masa studi yang mayoritas berada pada **4 tahun dan 4,5 tahun**. *Cluster* ini berasal dari mahasiswa yang mayoritas berasal dari **Tangerang, Bekasi dan Jakarta**. *Cluster* ini juga didominasi oleh mahasiswa yang mengambil **UKM Seni**. *Cluster* ini merupakan *cluster* yang dimana anggotanya mayoritas sudah mencoba untuk **hidup mandiri tanpa bergantung pada orangtua.**
1	*Cluster* ini memiliki anggota terbanyak kedua dengan **3.141 anggota**, didominasi dengan **pria daripada wanita**. Mayoritas anggota *cluster* ini lulus dengan **masa studi 5 tahun - 5,5 tahun.** Anggota dari *cluster* ini diisi oleh mahasiswa yang mayoritas **tinggal dengan orangtua** dan dibiayai oleh **orangtua semasa berkuliah.** Mahasiswa di *cluster* 1 juga kebanyakan **tidak bekerja** saat berkuliah, tetapi mengikuti **organisasi mahasiswa.** *Cluster* 1 kebanyakan adalah masyarakat yang **mendaftar kuliah pada tahun 2009**. Anggota *cluster* 1 berdomisili di daerah yang bervariasi, **didominasi oleh bogor, kemudian diikuti dengan Jakarta dan Bekasi, dan sisanya di Tangerang, Serang dan Karawang**. Anggota *cluster* 1 mayoritas adalah mahasiswa yang mengambil **fakultas FTI dan UKM kewirausahaan.** Berdasarkan karakteristiknya, *cluster* ini mayoritas anggotanya **berasal dari keluarga yang cukup mampu dan aktif berorganisasi.**
2	*Cluster* ini memiliki **2.251 anggota**, dimana **anggota pria mendominasi daripada wanita.** *Cluster* ini adalah *cluster* yang memiliki jangka waktu kuliah yang paling buruk dibanding *cluster* lain, dimana mayoritas anggota *cluster* ini lulus dengan **masa studi 7 tahun, diikuti dengan 5,5 - 6 tahun.** *Cluster* 2 ini juga **kebanyakan tidak bekerja dan tidak mengikuti organisasi mahasiswa** , serta **tidak mengikuti UKM** saat berkuliah. Namun,

	anggota *cluster* 2 ini **tidak begitu diketahui status pembiayaan kuliah dan status tinggalnya saat masih berkuliah**. Untuk domisili, *cluster* 2 didominasi oleh **mahasiswa yang tinggal di Karawang**. Anggota *cluster* 2 kebanyakan **memulai kuliah di tahun 2008,** dengan **fakultas FISIP** sebagai dominannya. Kesimpulan yang dapat diraih adalah *cluster* 2 merupakan ***cluster* yang anggotanya mayoritas pasif dalam menjalani masa studinya**.
3	*Cluster* ini memiliki anggota paling sedikit, yakni hanya **1.951 anggota**, dengan **anggota wanita yang lebih banyak dari pria**. *Cluster* 3 memiliki perbedaan yang cukup berseberangan dengan *cluster* 2, dimana *cluster* 3 **kebanyakan lulus dengan masa studi 3,5 tahun**. Untuk biaya kuliah , *cluster* 3 **didominasi oleh orang tua, sedangkan untuk status tinggal didominasi oleh kos.** Kemudian, *cluster* 3 juga mayoritas **sudah bekerja, mengikuti organisasi kemahasiswaan**, dan juga **mengikuti UKM yang didominasi olahraga** pada saat berkuliah. Lalu, mayoritas anggota *cluster* ini adalah mahasiswa yang **mendaftar kuliah pada tahun 2008**. Dapat dikatakan bahwa *cluster* 3 adalah ***cluster* yang cukup aktif pada masa berkuliah.**

Setelah melakukan *clustering* dan menganalisa hasil *clustering* yang ada, dapat ditarik sebuah kesimpulan bahwa *cluster 2* adalah kelompok mahasiswa yang perlu diberikan sebuah perhatian lebih, dimana kelompok mahasiswa tersebut mayoritas lulus dengan perhitungan yang sangat terlambat, yakni 7 tahun. Kelompok mahasiswa ini juga mayoritas tidak mengikuti kegiatan apapun diluar kampus, seperti bekerja, berorganisasi dan juga UKM. Padahal, komunitas merupakan sebuah kesempatan untuk mahasiswa membuat koneksi sejak dini yang membuat mahasiswa tetap bertahan dan mendapat Indeks Prestasi yang tinggi[15].

III. Penutup

i. Kesimpulan

Pada penelitian ini, telah dilakukan *Clustering* dengan algoritma *k-modes* yang dilakukan pada mahasiswa di suatu universitas yang sama. Penelitian ini diharapkan dapat membantu dosen pembimbing akademik untuk mengevaluasi dan menentukan kelompok mahasiswa seperti apa yang membutuhkan bimbingan akademik yang lebih intens guna dapat menyelesaikan masa studi dengan waktu yang lebih optimal.

Untuk mencari jumlah *cluster* yang optimal, digunakan 2 *evaluation metrics,* yaitu *Elbow Method* dan *Calinski-Harabasz Score*. Dari kedua metode tersebut, didapatkan hasil berupa k = 4. *Clustering* dilakukan dengan menggunakan metode 'Cao' untuk mencari titik inisiasi awal dan menghasilkan 4 buah *cluster,* yaitu *cluster* 0 dengan 3.931 anggota, *cluster* 1 dengan 3.141 anggota, *cluster 2* dengan 2.251 anggota, *cluster* 3 dengan 1.951 anggota.

Berdasarkan karakteristik keempat *cluster* yang ada, dapat disimpulkan bahwa *cluster* 2 adalah kelompok mahasiswa yang perlu diperhatikan lebih oleh dosen pembimbing, sebab mayoritas kelompok mahasiswa ini lulus dengan predikat "sangat terlambat". Hal ini tentunya sangat disayangkan, sebab dengan semakin lama kita menjalani masa studi, lebih lama juga kita membuang waktu dan uang yang seharusnya bisa kita gunakan untuk berkontribusi dalam kehidupan bermasyarakat.

ii. Saran

Saran yang dapat kami berikan dalam penelitian ini, yakni dengan menambahkan atau mencari referensi tambahan yang lebih spesifik, seperti nilai mahasiswa di semester awal (1-4) perkuliahan, jumlah absensi yang dilakukan, dan lain sebagainya. Kemudian, dapat juga dilakukan beberapa modifikasi pada bagian *Clustering*, dimana penentuan inisiasi titik awal dapat diganti dengan beberapa metode, seperti metode Huang yang menggunakan ukuran ketidaksamaan (*dissimilarity measure*) atau menginisiasi titik awal secara acak ataupun random tanpa metode apapun. Sebab, dengan titik awal yang berbeda, terdapat kemungkinan *cluster* yang terbentuk juga berbeda.

Daftar Pustaka

[1] Rahmat Hidayat, Dede (2021). “Perencanaan Akademis Mahasiswa”. Pelatihan Dosen PA Korpis Wilayah III 1-3 April 2021

[2] Huang, Z.X.(1997) "Clustering Large Data Sets with Mixed Numerical and Categorical Values". Asia Knowledge Discovery and Data Mining Conference, Singapore, World Scientific,pp. 21-34.

[3] V. Luckerson, "Time," 2013. (Online). Available at:http://business.time.com/2013/01/10/the-myth-of-the-4-year-college-degree/.(Diakses pada tanggal 3 September 2022)

[4] D. E. B. Monfatt, "Adult Undergraduates: Exploring Factors Essential to Success and Persistence toward Educational Goals," in Midwest Research-to-Practice Conference in Adult, Continuing, Community and Extension Education, Lindenwood University, St.Charles, MO, 2011.

[5] Kamus Besar Bahasa Indonesia.(Online). Available at https://kbbi.web.id/mahasiswa.(Diakses pada tanggal 4 September 2022)

[6] Badan Pemerintah Keuangan Republik Indonesia (BPK Republik Indonesia) (Online). Available at :https://peraturan.bpk.go.id/Home/Details/60869
(Diakses pada tanggal 4 September 2022)

[7] Davies, D. L.; Bouldin, D. W. "A Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence (2): 224, 1979.

[8] Fitria Febriani, Anita et.al. (2018). K-Means Clustering dengan Metode Elbow untuk Pengelompokkan Kabupaten dan Kota di Jawa Timur berdasarkan Indeks Kemiskinan. Available at https://karyailmiah.unipasby.ac.id/wp-content/uploads/2019/04/K-Means-Artikel.pdf.

[9] T. Caliński & J Harabasz (1974) A dendrite method for cluster analysis, Communications in Statistics - Theory and Methods, 3:1, 1-27

[10] Ronaghan, Stacey (Online). Data - AI - Analytics Reference Literature - Data Understanding . Available at https://ibm-cloud-architecture.github.io/refarch-data-ai-analytics/preparation/data-understanding/ (Diakses pada tanggal 5 September 2022)

[11] S. LakshmiMphil et al. (2018). An Overview Study on Data Cleaning, Its Types and Its Methods for Data Mining. International Journal of Pure and Applied Mathematics , Vol. 119 No. 12

[12] Yadav, Dinesh. (Online). Categorical Encoding using Label - Encoding and One- Hot Encoder. Available at: https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd.
(Diakses pada tanggal 7 September 2022)

[13] Seif, George (Online). The 5 Clustering Algorithms Data Scientists Need to Know. Available at https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 (Diakses pada tanggal 5 September 2022)

[14] Fuyuan, Cao et al. 2009 . A New Initialization Method for Categorical Data Clustering. International Journal of Expert System with Applications.

[15] Rohli, R.V. and Rogge, R.A. (2012, February). An empirical study of potential for geography in university living-learning communities in the United States. Journal of Geography in Higher Education, 36(2), 81-95.