Volume 4, Nomor 2, Desember 2023, hlm 234-246
INFOTECH: Jurnal Informatika & Teknologi
p ISSN 2722-9378 | e ISSN 2722-9386
http://jurnal.sttmcileungsi.ac.id/index.php/infotech

Analisis cluster data latar belakang ekonomi mahasiswa untuk rekomendasi penentuan uang kuliah tunggal dengan model K-Modes

Cluster analysis of student's economic background data for recommendation to determining single tuition using K-Modes model

Adi Wirawan*, Daru Prasetyawan

*Pusat Teknologi Informasi dan Pangkalan Data, UIN Sunan kalijaga Yogyakarta, Jl. Marsda Adisudipto Yogyakarta, Indonesia

INFORMASI ARTIKEL

Article History:
Submission: 13-11-2023
Revised: 28-11-2023
Accepted: 01-12-2023

Kata Kunci:
Clustering; data kategorik; k-modes; latar belakang ekonomi; ukt

Keywords:
Clustering; categorical data; k-modes; socio-economic background; tuition

*** Korespondensi:**
Adi Wirawan
adi.wirawan@uin-suka.ac.id

ABSTRAK

Latar belakang ekonomi mahasiswa merupakan salah satu faktor yang berpengaruh dalam keberhasilan seorang mahasiswa dalam menyelesaikan kuliahnya. Pengelompokan mahasiswa berdasarkan latar belakang ekonomi dapat digunakan untuk identifikasi kemampuan ekonomi mahasiswa. Penelitian ini bertujuan untuk pengelompokan data latar belakang ekonomi mahasiswa berdasarkan atribut-atribut di dalamnya, seperti penghasilan per kapita, status kepemilikan rumah, penggunaan daya listrik, jumlah mobil, jumlah motor, biaya pulsa dan internet, serta jaminan biaya pendidikan menggunakan algoritma k-Modes untuk rekomendasi dalam penentuan Uang Kuliah Tunggal (UKT). K-Modes digunakan dalam clustering ini karena k-Modes dapat menangani data kategorik dengan baik. Dalam proses *clustering* dengan menggunakan k-Modes, metode Elbow digunakan untuk mencari jumlah cluster yang paling optimal. Dari proses clustering, data latar belakang mahasiswa dikelompokan menjadi 3 *cluster*. *Cluster* pertama memiliki karakteristik dengan latar belakang ekonomi relatif rendah, cluster kedua memiliki karakteristik latar belakang ekonomi sedang, dan cluster ketiga memiliki karakteristik latar belakang ekonomi tinggi. Hasil dari dari proses analisis cluster tersebut selanjutnya digunakan sebagai rekomendasi dalam penentuan UKT.

ABSTRACT

A student's economic background is one of the factors that influences a student's success in completing their studies. Grouping students based on economic background can be used to identify students' economic abilities. This research aims to group student economic background data based on attributes in it, such as income per capita, home ownership status, electricity usage, number of cars, number of motorbikes, credit and internet costs, as well as guaranteed education costs using the k-Modes algorithm for recommendations in determining the Single Tuition Fee (UKT). K-Modes is used in this clustering because k-Modes can handle categorical data well. In the clustering process using k-Modes, the Elbow method is used to find the most optimal number of clusters. From the clustering process, student background data is grouped into 3 clusters. The first cluster is characterized by a relatively low economic background, the second cluster is characterized by a medium economic background, and the third cluster is characterized by a high economic background. The results of the cluster analysis process are then used as recommendations in determining UKT.

1. PENDAHULUAN

Latar belakang sosial ekonomi keluarga merupakan salah satu faktor yang mempergaruhi kelancaran proses belajar mahasiswa di perguruan tinggi. Kondisi ekonomi keluarga yang kurang mampu sering menjadi kendala bagi seorang mahasiswa dalam menyelesaikan studinya. Keluarga dengan kecukupan ekonomi dapat memberikan lingkungan materil yang lebih luas, sehingga memperoleh kesempatan di berbagai bidang [1]. Perhatian yang lebih baik terhadap pemenuhan kebutuhan untuk masa depan anak-anaknya biasanya dilakukan oleh keluarga dengan status sosial ekonomi yang baik [2]. Sedangkan keluarga dengan kondisi ekonomi kurang mampu, akan lebih mengutamakan untuk memenuhi kebutuhan pokok, dan kurang memberikan perhatian dalam peningkatan pendidikan anak [3]. Selain menyelesaikan tugas sebagai mahasiswa, terkadang mereka juga bekerja paruh waktu untuk memenuhi kebutuhan sehari-hari. Sehingga mereka harus pandai dalam membagi waktu.

Pengelompokan mahasiswa baru berguna untuk membantu mengidentifikasi kemampuan ekonomi mahasiswa dalam upaya mendukung kelancaran proses belajar mahasiswa di perguruan tinggi. Analisis *cluster* dapat menemukan pola dan karakteristik latar belakang ekonomi mahasiswa sebagai bahan pertimbangan dalam menentukan kebijakan yang terkait dengan pengembangan mahasiswa, seperti penentuan UKT (Uang Kuliah Tunggal), pemberian bantuan beasiswa, dan bantuan konseling. UKT merupakan sistem pembayaran kuliah yang dilakukan oleh Perguruan Tinggi Negeri dengan menyesuaikan kondisi perekonomian mahasiswa. UKT terdiri dari beberapa kelompok yang ditentukan berdasarkan kemampuan ekonomi mahasiswa, orang tua, dan pihak lain. Pengelompokan UKT dilakukan dengan mempertimbangkan banyak variabel seperti penghasilan orang tua/wali, hutang keluarga, tanggungan keluarga, luas tanah yang dimiliki, tagihan listrik, tagihan air, tagihan internet, kepemilikan motor/mobil, dan sebagainya. Namun dalam kenyataannya, pengelompokan UKT hanya didasarkan pada gaji orang tua/wali saja. Hal ini dikarenakan sulitnya untuk mengelompokan data dengan dimensi yang banyak dan tipe data yang berbeda.

Perkembangan teknologi khususnya di bidang kecerdasan buatan dan pembelajaran mesin telah banyak dimanfaatkan dalam berbagai bidang. Kecerdasan Buatan (*Artificial Intelligence*) dapat memberikan harapan terbaik di banyak sektor aplikasi di seluruh bidang jika dimanfaatkan dengan tepat [4]. Pembelajaran Mesin (*Machine Learning*) adalah salah satu cabang dari kecerdasan buatan yang fokus pada pengembangan teknik dan algoritma yang memungkinkan suatu komputer dapat belajar dari data dan pengalaman, serta meningkatkan kinerjanya, tanpa harus diprogram secara eksplisit. Pembelajaran mesin melibatkan penggunaan algoritma komputer untuk menganalisis data, mengenali pola, dan membuat keputusan atau prediksi berdasarkan data yang diberikan. Proses pembelajaran dilakukan dengan pengamatan terhadap sekumpulan data dan melatih mesin menggunakan data tersebut, sehingga mesin tersebut dapat belajar secara mandiri tanpa campur tangan manusia [5].

Pembelajaran mesin terbagi menjadi 3 jenis, yaitu pembelajaran terawasi (*supervised learning*), pembelajaran tak terawasi (*unsupervised learning*), dan pembelajaran penguatan (*reinforcement learning*). Pembelajaran terawasi merupakan pembelajaran mesin yang mampu menghasilkan pola dan hipotesis umum dengan menggunakan *instance* yang disediakan secara eksternal untuk memprediksi *instance* di masa depan [6]. Pada pembelajaran terawasi, mesin atau komputer membuat keputusan atau prediksi dengan melakukan proses pembelajaran menggunakan data pelatihan yang sudah memiliki label. Tujuan utama pembelajaran mesin adalah untuk membuat model atau algoritma yang dapat memetakan data input ke data output yang sesuai. Pembelajaran mesin tak terawasi merupakan jenis pembelajaran mesin yang dapat menemukan pola dalam kumpulan data yang tidak berlabel [7]. Pembelajaran tak terawasi bertujuan untuk mengekstrak informasi yang berguna dari data, seperti pengelompokan (*clustering*) atau reduksi dimensi (*dimensionality reduction*). Pembelajaran penguatan adalah paradigma dalam pembelajaran mesin yang memungkinkan sebuah agen belajar untuk mengambil tindakan dalam lingkungan tertentu untuk mencapai tujuan tertentu. Pembelajaran penguatan melakukan interaksi dengan lingkungan dengan memperbaiki kegagalan yang dialami dan memaksimalkan imbalan yang diterima [8]. Pembelajaran penguatan dapat menghasilkan kebijakan dengan pengetahuan yang diperoleh dari proses *trial-and-error* [9]. Dalam

pembelajaran penguatan, agen berinteraksi dengan lingkungannya dan belajar dari pengalaman, menerima umpan balik positif atau negatif dalam bentuk imbalan dari hasil dari tindakan yang diambil. Pembelajaran perkuatan digunakan untuk mempelajari strategi atau kebijakan (*policy*) yang memaksimalkan akumulasi imbalan seiring berjalannya waktu.

Salah satu jenis pembelajaran mesin adalah pembelajaran tak terawasi (*unsupervised learning*), termasuk di dalamnya adalah *clustering*. Jenis pembelajaran mesin tersebut sering digunakan untuk pengelompokan data dengan banyak variabel. *Clustering* memisahkan sekumpulan variabel hasil pengukuran atau perhitungan ke dalam kelompok yang homogen [10]. Analisis *cluster* atau sering dikenal dengan *clustering* merupakan teknik untuk membagi sekumpulan data menjadi menjadi beberapa cluster atau kelompok yang terdiri dari objek-objek yang serupa. *Clustering* adalah sebuah konsep untuk menentukan pola melalui pemetaan dan analisis data [11]. Clustering adalah metode statistik multivariat modern dalam mempelajari kesamaan sampel yang berbeda, dengan mengelompokkan sekumpulan objek ke dalam kelas-kelas atau *cluster* sedemikian rupa sehingga objek-objek dalam suatu cluster memiliki kemiripan antara satu dengan yang lain. Kemiripan antar objek dalam satu grup dibuat setinggi-tingginya, dan kemiripan objek antar grup yang berbeda dibuat seminimal mungkin [12]. Clustering digunakan untuk mengelompokkan data atau objek menjadi kelompok-kelompok yang serupa berdasarkan karakteristik atau atribut yang dimiliki oleh objek tersebut. Pengelompokan tersebut mengacu pada pemecahan sekumpulan data menjadi kelompok-kelompok menurut kriteria yang sesuai dengan mengasosiasikan sampel data melalui kemiripan atau ketidakmiripan [13].

Dalam proses pengelompokan data, banyak algoritma *clustering* yang digunakan. Algoritma pengelompokan yang terkenal adalah k-Means. Tetapi k-Means hanya dapat digunakan pada data numerik saja. Sedangkan dalam dunia nyata, terdapat banyak data kategorik yang harus dipertimbangkan dalam analisis. Oleh karena itu, pada tahun 1998, Huang memperkenalkan metode *clustering* baru yang dikembangkan dari metode k-Means untuk menangani data kategorik yang dikenal dengan k-Modes [14]. Modifikasi terjadi dalam menentukan centroid dengan mengganti jarak (*distance*) menjadi *dissimilarity measure* dan mengganti *means* menjadi *modes* atau nilai atribut yang paling sering muncul dari suatu atribut. Untuk menentukan *mode*, k-Modes menggunakan frekuensi atau nilai yang muncul sering muncul.

K-Modes merupakan algoritma *clustering* yang sering digunakan untuk mengelompokan data kategorik karena mudah diimplementasikan dan efisien untuk menangani data dalam jumlah besar [15]. Algoritma k-Modes menjamin konvergensi, artinya algoritma tersebut dapat memberikan hasil yang pasti. Algoritma k-Modes dapat menemukan *centroid* yang merupakan pola valid dan benar-benar mewakili sebuah *cluster* [16]. K-Modes sangat mudah dipahami dan diimplementasikan karena tidak menggunakan konstruksi matematika tingkat lanjut. Selain itu, k-Modes dapat digunakan pada berbagai domain data, artinya selama kumpulan data tersebut memiliki atribut kategorik, algoritma k-Modes dapat digunakan untuk mempartisi kumpulan data ke dalam *cluster* yang berbeda.

Clustering menggunakan metode k-Modes telah digunakan untuk pengelompokan jenis masakan berdasarkan bahan-bahan yang digunakan. Berdasarkan metode *Elbow*, diperoleh jumlah cluster terbaik sebanyak 4 dan 8 [17]. Dalam bidang keamanan jaringan, k-Modes juga digunakan dalam mendeteksi serangan atau intrusi pada jaringan internet dengan membaginya menjadi 2 *cluster*, yaitu normal dan abnormal [12]. K-Modes juga diterapkan dalam pengelompokan karakteristik calon Tenaga Kerja Indonesia dan terbentuk 2 *cluster* [18].

Dalam pengelompokan mahasiswa, k-Modes digunakan dalam pengelompokan tingkat stres mahasiswa [19]. K-Modes juga dapat digunakan untuk mengukur tingkat kepuasan mahasiswa dalam proses pembelajaran daring dan menghasilkan 4 *cluster*, yaitu *cluster* sangat puas, puas, tidak puas, dan sangat tidak puas [20]. Implementasi k-Modes dalam pengelompokan data mahasiswa juga dapat digunakan sebagai strategi pemasaran sebuah perguruan tinggi, dengan melakukan analisis *cluster* terhadap data asal sekolah mahasiswa [21].

Berdasarkan latar belakang dan kajian pustaka terhadap penelitian-penelitian sebelumnya, penelitian ini melakukan analisis *cluster* terhadap data latar belakang ekonomi mahasiswa menggunakan algoritma k-Modes dengan tujuan untuk memperoleh pola dan karakteristik latar belakang ekonomi mahasiswa. Untuk memperoleh hasil terbaik, metode *Elbow* digunakan dalam menentukan nilai k yang optimal. Dengan memanfaatkan teknologi pembelajaran mesin, dalam

hal ini *clustering*, pengelompokan data latar belakang ekonomi mahasiswa dapat dengan mudah dilakukan dan akan memperoleh hasil yang lebih komprehensif.

2. METODE

2.1 Format nama penulis.

Dataset yang digunakan untuk analisis cluster ini adalah data latar belakang ekonomi mahasiswa baru UIN Sunan Kalijaga Yogyakarta tahun akademik 2023/2024 Fakultas Sains dan Teknologi ditambah dengan program studi Pendidikan MIPA dengan jumlah data sebanyak 736 data. Data tersebut diperoleh dari data yang diisikan oleh mahasiswa baru yang dinyatakan diterima melalui jalur penerimaan mahasiswa baru. Mahasiswa yang dinyatakan diterima pada program studi tertentu diminta untuk mengisikan data latar belakang ekonomi pada aplikasi data profil yang dapat diakses melalui laman *dataprofil.uin-suka.ac.id*. Selanjutnya data tersebut divalidasi oleh petugas atau operator aplikasi dengan mencocokan data yang diisikan dengan lampiran dokumen yang juga diunggah melalui aplikasi tersebut.

Data latar belakang ekonomi mahasiswa baru memiliki 7 atribut atau fitur yang digunakan dalam analisis *cluster*, yaitu penghasilan keluarga per kapita, status kepemilikan rumah, penggunaan daya listrik, jumlah mobil yang dimiliki, jumlah sepeda motor yang dimiliki, biaya pulsa dan internet, serta kepemilikan jaminan pendidikan dan jaminan sosial. Deskripsi data latar belakang ekonomi dapat dilihat pada Tabel 1.

Tabel 1. Deskripsi data latar belakang ekonomi

column	count	unique	top	freq
penghasilan_per_kapita	736	5	300.001 - 500.000	181
kepemilikan_rumah	736	2	Sendiri	671
daya_listrik	736	3	900	373
jumlah_mobil	736	2	0	586
jumlah_motor	736	4	2	335
biaya_pulsa_internet	736	4	0 - 50.000	219
jaminan_pendidikan	736	2	Tidak ada	658

2.2 Pemilihan Nilai k.

K-Modes pengelompokan data menggunakan metrik pencocokan yang mengukur perbedaan dari dua titik data kategoris. Seperti halnya k-Means, k-Modes tidak mengetahui jumlah *cluster* yang pasti untuk pengelompokan data. Oleh karena itu, untuk menentukan nilai k terbaik perlu mencoba satu per satu nilai k yang berbeda. Metode *Elbow* merupakan metode yang sering digunakan untuk menentukan nilai k yang optimal [22]. Metode *Elbow* digambarkan sebagai grafik nilai *cost* untuk setiap nilai k yang membentuk seperti siku [19]. *Cost* merupakan jumlah dari semua ketidaksamaan antar cluster atau *within-cluster-difference* (WCD) seperti pada persamaan 1 [23].

$$wcd = \sum_{j=1}^{k} \sum_{i=1}^{m} d_1(x_1, y_c) \quad (1)$$

Dimana k adalah jumlah *cluster*, m ada jumlah data di dalam setiap *cluster*, c merupakan *centroid cluster*, dan d_1 merupakan ukuran ketidaksamaan sederhana (*simple dissimilarity measure*).

2.3 Proses *clustering* dengan K-Modes.

Setelah nilai k sudah ditentukan, selanjutnya nilai k tersebut digunakan dalam proses *clustering* dengan K-Modes. Di dalam k-Modes, penghitungan jarak di antara 2 objek dihitung dengan ukuran ketidaksamaan sederhana (*simple dissimilarity measure*) yang diformulasikan pada persamaan 2 dan persamaan 3 [14].

$$d_1(X, Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \quad (2)$$

dimana:

$$\delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases} \quad (3)$$

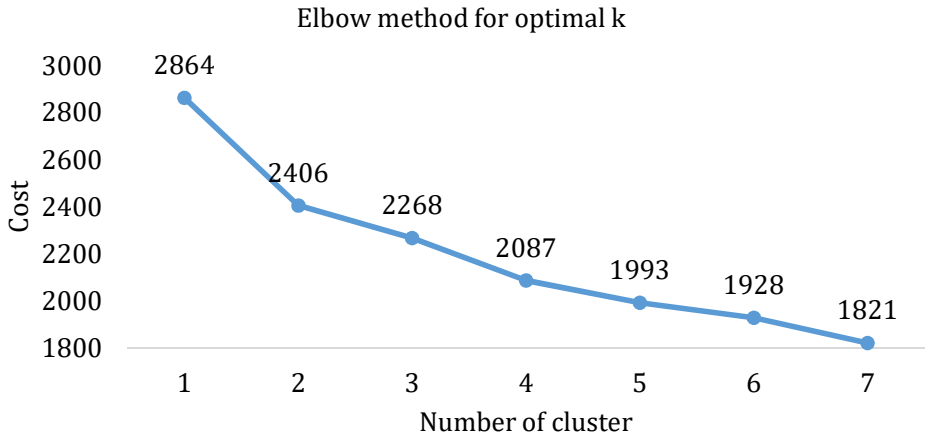
dengan x_i merupakan nilai data ke-i dari data X, y_i adalah nilai data ke-i dari data Y, dan n adalah jumlah data observasi.

Langkah-langkah *clustering* dengan k-Modes sebagai berikut [14]:

1. Pilih mode awal setiap k
2. Alokasikan objek data ke cluster terdekat berdasarkan ukuran ketidaksamaan sederhana (*simple dissimilarity measure*). Perbarui *mode cluster* setiap pengalokasian objek.
3. Selanjutnya, periksa nilai ketidaksamaan (*dissimilarity*) antara setiap objek dan *mode*. Jika suatu data lebih dekat dengan *mode* pada *cluster* lain, pindahkan objek ke *cluster* yang sesuai dan perbarui *mode cluster.*
4. Ulangi langkah 3 hingga tidak ada objek data yang berpindah cluster.

3. HASIL DAN PEMBAHASAN

Proses analisis *cluster* dalam penelitian ini menggunakan bahasa pemrograman Python. Algoritma k-Modes tidak dapat menentukan jumlah *cluster* yang terbaik, oleh karena itu setiap jumlah *cluster* harus dicoba satu per satu. Jumlah *cluster* optimal yang akan digunakan dalam analisis *cluster* dapat diketahui dengan menggunakan metode Elbow. Di dalam analisis cluster menggunakan k-Modes, metode Elbow dibentuk melalui perhitungan *cost* atau WCD seperti pada persamaan 1. Nilai k yang membentuk siku merupakan k yang optimal. Gambar 1 Menunjukan grafik nilai *cost* pada setiap kandidat yang akan menjadi nilai k.



Gambar 1. Grafik nilai *cost* pada setiap kandidat yang akan menjadi nilai k

Dari Gambar 1 bahwa grafik *cost* membentuk siku pada k=3, sehingga jumlah *cluster* yang akan di gunakan dalam analisis *cluster* adalah 3. Setelah jumlah cluster telah ditentukan, selanjutnya proses *clustering* dapat dilakukan. Tabel 2 menunjukan data latar belakang ekonomi mahasiswa.

Tabel 2. Sampel data latar belakang ekonomi mahasiswa

id	penghasilan_per_kapita	Kepemilikan_rumah	daya_listrik	Jumlah_mobil	Jumlah_motor	biaya_pulsa_internet	jaminan_pendidikan
1	300.001 - 500.000	Sendiri	900	0	0	0 - 50.000	Ada
2	0 - 300.000	Sendiri	450	0	2	50.001 - 100.000	Tidak ada
3	di atas 1.500.000	Sendiri	900	0	2	0 - 50.000	Tidak ada
4	1.000.001 - 1.500.000	Sendiri	450	0	2	di atas 200.000	Tidak ada
5	300.001 - 500.000	Sewa/Kotrak	900	0	1	0 - 50.000	Tidak ada
...	...	...	...	...	...	...	...
732	di atas 1.500.000	Sendiri	1300 ke atas	1	2	0 - 50.000	Tidak ada

id	penghasilan_ per_kapita	Kepemilikan _rumah	daya_ listrik	Jumlah _mobil	Jumlah _motor	biaya_pulsa _internet	jaminan_ pendidikan
733	1.000.001 - 1.500.000	Sendiri	450	0	1	100.001 - 200.000	Tidak ada
734	1.000.001 - 1.500.000	Sendiri	1300 ke atas	0	2	di atas 200.000	Tidak ada
735	500.001 - 1.000.000	Sendiri	900	0	lebih dari 2	100.001 - 200.000	Tidak ada
736	0 - 300.000	Sendiri	450	0	1	50.001 - 100.000	Tidak ada

Setelah nilai k ditentukan, Langkah kedua adalah memilih *leader* untuk setiap *cluster* secara acak. Sebagai contoh data ke-2 sebagai *leader cluster* C1, data ke-5 sebagai *leader* C2, dan data ke-732 sebagai *leader* C3 seperti pada Tabel 3.

Tabel 3. Inisialisasi cluster awal

Cluster	penghasilan_ per_kapita	Kepemilikan _rumah	daya_ listrik	Jumlah _mobil	Jumlah _motor	biaya_pulsa _internet	jaminan_ pendidikan
2	0 - 300.000	Sendiri	450	0	2	50.001 - 100.000	Tidak ada
5	300.001 - 500.000	Sewa/Kotrak	900	0	1	0 - 50.000	Tidak ada
732	di atas 1.500.000	Sendiri	1300 ke atas	1	2	0 - 50.000	Tidak ada

Langkah ketiga adalah membandingkan setiap data observasi dengan semua *leader/cluster* dan dihitung ketidaksamaannya. Misalnya:

$d_1(X_1,C_1) = 1+0+1+0+1+1+1 =5$

$d_1(X_1,C_2) = 0+1+0+0+1+0+1 =3$

$d_1(X_1,C_3) = 1+0+1+1+1+0+1 =5$

Kemudian masukan data observasi ke dalam *cluster* dengan nilai ketidaksamaan terkecil seperti pada Tabel 4.

Tabel 4. Matriks ketidaksamaan data observasi

id	Cluster C1	Cluster C2	Cluster C3	Cluster
1	5	**3**	5	C2
2	**0**	5	4	C1
3	3	3	**2**	C3
4	**2**	5	4	C1
5	5	**0**	5	C2
...	...	...	...	...
732	4	5	**0**	C3
733	**3**	4	5	C1
734	**3**	5	3	C1
735	**4**	4	5	C1
736	**1**	4	5	C1

Langkah keempat adalah menentukan nilai *mode*, yaitu nilai yang paling sering muncul. Hasil penentuan mode setiap cluster dapat pada Tabel 5.

Tabel 5. Mode setiap cluster

Mode	penghasilan_ per_kapita	Kepemilikan _rumah	daya_ listrik	Jumlah _mobil	Jumlah _motor	biaya_pulsa _internet	jaminan_ pendidikan
C1	0 - 300.000	Sewa/Kotrak	450	0	1	50.001 - 100.000	Ada
C2	300.001 - 500.000	Sendiri	900	0	2	0 - 50.000	Tidak ada

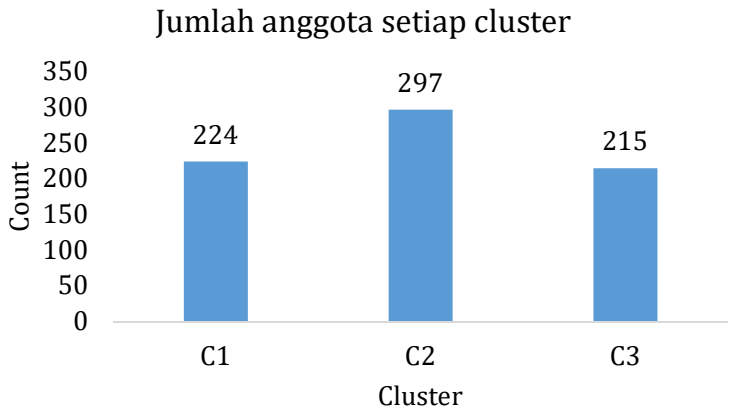
Mode	penghasilan_ per_kapita	Kepemilikan _rumah	daya_ listrik	Jumlah _mobil	Jumlah _motor	biaya_pulsa _internet	jaminan_ pendidikan
C3	di atas 1.500.000	Sendiri	1300 ke atas	1	Lebih dari 2	di atas 200.000	Tidak ada

Selanjutnya ulangi langkah ketiga dan keempat sehingga tidak ada lagi data observasi yang berpindah *cluster*. Setelah proses *clustering* selesai, diperoleh pengelompokan data pada *cluster* yang sesuai seperti pada Tabel 6.

Tabel 6. Hasil proses clustering data latar belakang ekonomi mahasiswa

id	penghasilan_ per_kapita	Kepemilikan _rumah	daya_ listrik	Jumlah _mobil	Jumlah _motor	biaya_pulsa _internet	jaminan_ pendidikan	Cluster
1	300.001 - 500.000	Sendiri	900	0	0	0 - 50.000	Ada	C2
2	0 - 300.000	Sendiri	450	0	2	50.001 - 100.000	Tidak ada	C1
3	di atas 1.500.000	Sendiri	900	0	2	0 - 50.000	Tidak ada	C3
4	1.000.001 - 1.500.000	Sendiri	450	0	2	di atas 200.000	Tidak ada	C3
5	300.001 - 500.000	Sewa/Kotrak	900	0	1	0 - 50.000	Tidak ada	C1
...	...	...	...	...	...	...	...	...
732	di atas 1.500.000	Sendiri	1300 ke atas	1	2	0 - 50.000	Tidak ada	C3
733	1.000.001 - 1.500.000	Sendiri	450	0	1	100.001 - 200.000	Tidak ada	C1
734	1.000.001 - 1.500.000	Sendiri	1300 ke atas	0	2	di atas 200.000	Tidak ada	C3
735	500.001 - 1.000.000	Sendiri	900	0	lebih dari 2	100.001 - 200.000	Tidak ada	C2
736	0 - 300.000	Sendiri	450	0	1	50.001 - 100.000	Tidak ada	C1

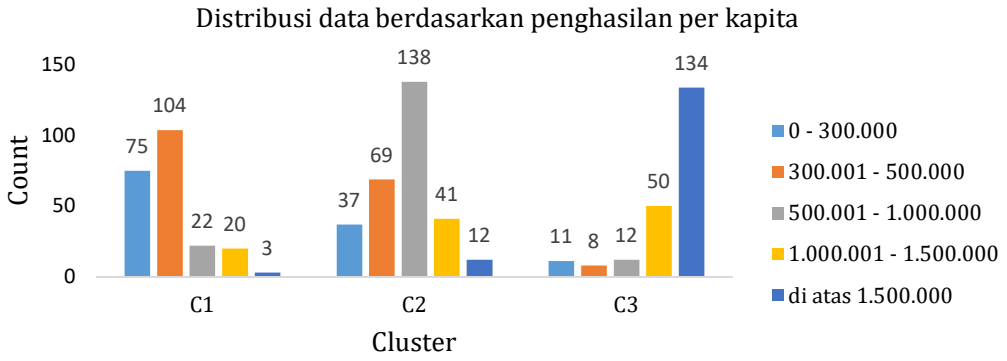
Dari hasil proses *clustering*, data latar belakang ekonomi mahasiswa dikelompokan menjadi 3 cluster, yaitu C1, C2, dan C3. Cluster C1 memiliki anggota sebanyak 224 data, *cluster* C2 memiliki anggota sebanyak 297 data, dan *cluster* C3 memiliki anggota sebanyak 215 data. Jumlah anggota pada masing-masing *cluster* disajikan pada Gambar 2.



Gambar 2. Jumlah anggota pada masing-masing cluster

Setiap kategori pada atribut penghasilan per kapita tersebar pada semua *cluster*. Pada *cluster* C1 kategori penghasilan per kapita yang paling banyak adalah "0 – 300.000" dan "300.001 – 500.000". Data dengan kategori penghasilan per kapita "500.001 – 1.000.000" menjadi anggota terbanyak di *cluster* C2. Sedangkan *cluster* C3 lebih didominasi oleh data dengan kategori

penghasilan per kapita "di atas 1.500.000". Distribusi data di dalam *cluster* berdasarkan penghasilan per kapita dapat dilihat pada Gambar 3.



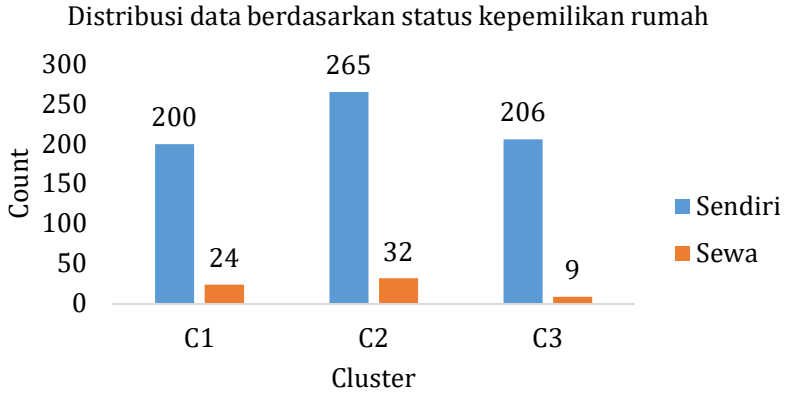
Gambar 3. Distribusi data di dalam *cluster* berdasarkan penghasilan per kapita

Pada kategori penghasilan per kapita "0 – 300.000" dan "300.001 – 500.000", persentase terbanyak berada pada *cluster* C1, sebesar 60,98% dan 57,46%. Untuk kategori "500.001 – 1.000.000" data terbanyak berada pada *cluster* C2 (80,23%). Sedangkan pada kategori "10.000.001 – 1.500.001" dan kategori "di atas 1.500.000" data terbanyak berada pada *cluster* C3, yaitu sebesar 45,05% dan 89,93. Persentase sebaran data setiap kategori pada atribut penghasilan per kapita selengkapnya disajikan pada Tabel 7.

Tabel 7. Persentase sebaran data setiap kategori pada atribut penghasilan per kapita

Kategori	Cluster C1	Cluster C2	Cluster C3
0 - 300.000	**75 (60.98%)**	37 (30.08%)	11 (8.94%)
300.001 - 500.000	**104 (57.46%)**	69 (38.12%)	8 (4.42%)
500.001 - 1.000.000	22 (12.79%)	**138 (80.23%)**	12 (6.98%)
1.000.001 - 1.500.000	20 (18.02%)	41 (36.94%)	**50 (45.05%)**
di atas 1.500.000	3 (2.01%)	12 (8.05%)	**134 (89.93%)**

Berdasarkan jumlah data pada atribut status kepemilikan rumah, semua *cluster* didominasi oleh data dengan kategori "Sendiri". Hal ini terjadi karena jumlah data antar kategori yang tidak seimbang. Distribusi data berdasarkan atribut kepemilikan rumah dapat dilihat pada Gambar 4.



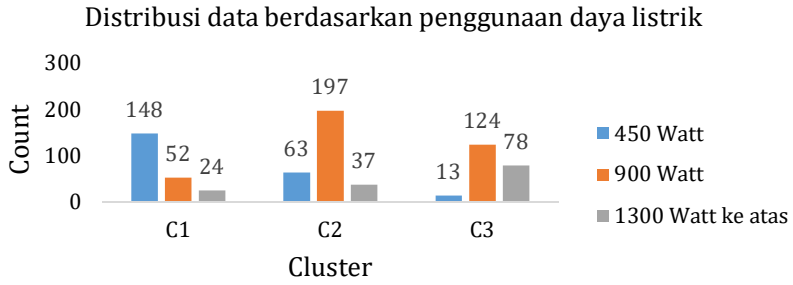
Gambar 4. Distribusi data berdasarkan atribut status kepemilikan rumah

Data dengan kategori kepemilikan rumah "Sendiri" tersebar merata di semua cluster, sedangkan data dengan kategori "Sewa/Kontrak" lebih banyak berada pada *cluster* C1 dan C2, dan hanya sedikit saja yang berada pada *cluster* C3. Persentase sebaran data setiap kategori pada atribut status kepemilikan rumah disajikan pada Tabel 8.

Tabel 8. Persentase sebaran data setiap kategori pada atribut status kepemilikan rumah

Kategori	Cluster C1	Cluster C2	Cluster C3
Sendiri	200 (29.81%)	265 (39.49%)	**206 (30.70%)**
Sewa/Kontrak	**24 (36.92%)**	**32 (49.23%)**	9 (13.85%)

Berdasarkan jumlah data pada atribut penggunaan daya listrik, distribusi data terbanyak pada *cluster* C1 adalah data dengan kategori penggunaan daya listrik "450 Watt". Jumlah data terbanyak pada *cluster* C2 adalah data dengan kategori "900 Watt". Sedangkan pada *cluster* C3, kategori "900 Watt" dan "1300 Watt" ke atas" lebih banyak dibandingkan dengan kategori "450 Watt". Distribusi data di dalam *cluster* berdasarkan atribut penggunaan daya listrik dapat dilihat pada Gambar 5.



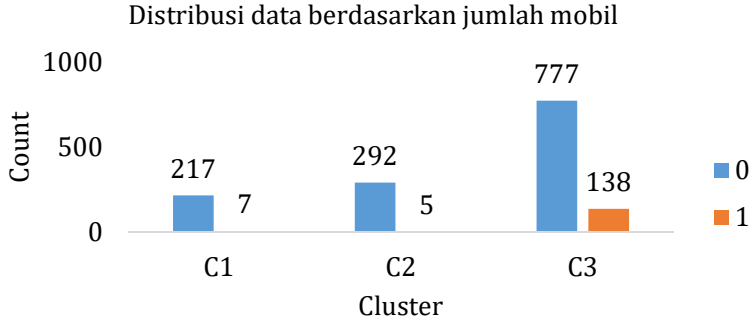
Gambar 5. Distribusi data di dalam *cluster* berdasarkan atribut penggunaan daya listrik

Berdasarkan persentase sebaran data setiap kategori pada atribut penggunaan daya listrik, jumlah data dengan kategori "450 Watt" terbanyak berada pada *cluster* C1. Jumlah data dengan kategori "900 Watt" paling banyak berada pada *cluster* C2, dan data dengan kategori "1300 Watt ke atas" berada pada *cluster* C3. Persentase sebaran data setiap kategori pada atribut penggunaan daya listrik disajikan pada Tabel 9.

Tabel 9. Persentase sebaran data setiap kategori pada atribut penggunaan daya listrik

Kategori	Cluster C1	Cluster C2	Cluster C3
450 Watt	**148 (66.07%)**	63 (28.13%)	13 (5.80%)
900 Watt	52 (13.94%)	**197 (52.82%)**	124 (33.24%)
1300 Watt ke atas	24 (17.27%)	37 (26.62%)	**78 (56.12%)**

Berdasarkan atribut jumlah mobil, kategori jumlah mobil "0" sangat dominan pada *cluster* C1 dan C2. Sedangkan pada *cluster* C3, kategori jumlah mobil "1" lebih banyak dari pada jumlah data dengan kategori "0". Distribusi data di dalam *cluster* berdasarkan atribut jumlah mobil dapat dilihat pada Gambar 6.



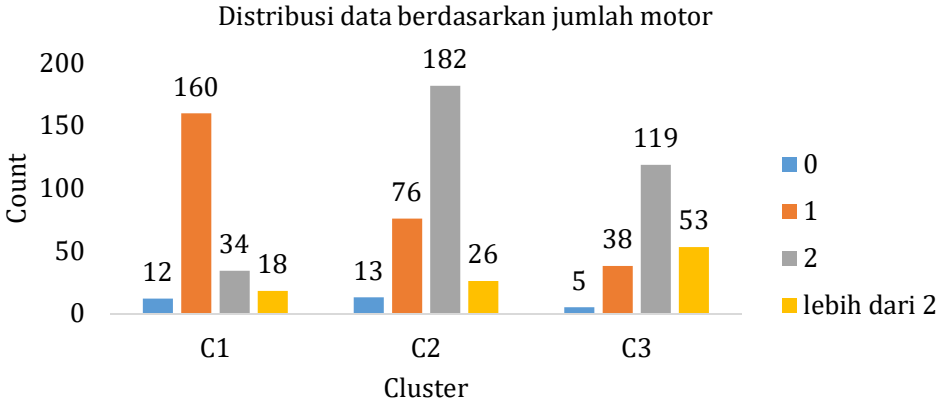
Gambar 6. Distribusi data di dalam cluster berdasarkan atribut jumlah mobil

Sebaran data setiap kategori pada atribut jumlah mobil terlihat bahwa data dengan kategori "0" lebih banyak berada pada *cluster* C1 (37,03%) dan C2 (49,83%), sedangkan kategori "1" lebih banyak berada pada *cluster* C3 (90%). Persentase sebaran data berdasarkan atribut jumlah mobil disajikan pada Tabel 10.

Tabel 10. Persentase sebaran data berdasarkan atribut jumlah mobil

Kategori	Cluster C1	Cluster C2	Cluster C3
0	**217 (37.03%)**	**292 (49.83%)**	77 (13.14%)
1	7 (4.67%)	5 (3.33%)	**138 (90%)**

Berdasarkan atribut jumlah motor, data dengan kategori jumlah motor "1" mendominasi cluster C1. Data dengan kategori jumlah motor "2" menjadi kategori dengan jumlah terbanyak pada *cluster* C2 dan C3. Distribusi data di dalam *cluster* berdasarkan atribut jumlah motor dapat dilihat pada Gambar 7.



Gambar 7. Distribusi data di dalam *cluster* berdasarkan atribut jumlah motor

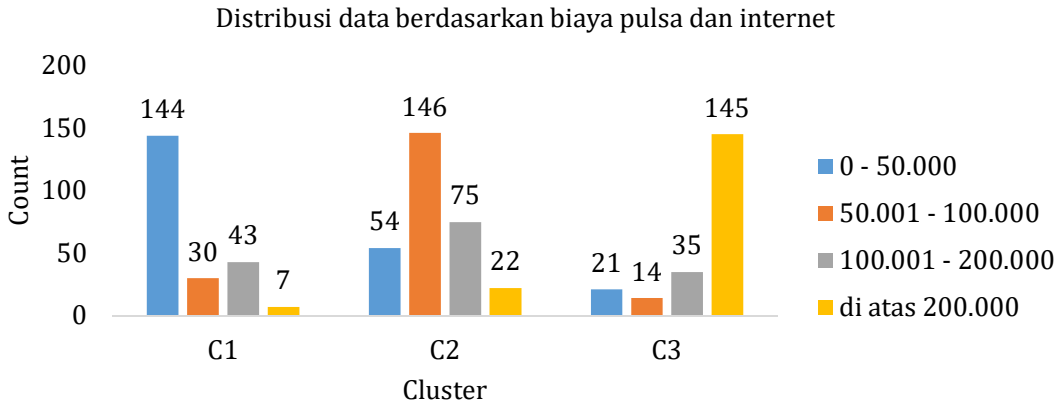
Berdasarkan sebaran data setiap kategori pada atribut jumlah motor, data dengan kategori "0" tersebar hampir seimbang di *cluster* C1 dan C1, tetapi hanya sedikit saja yang berada pada cluster C3. Sebaran data dengan kategori "1" terbanyak berada pada *cluster* C1 (58,18%). Data dengan kategori "2" paling banyak berada pada *cluster* C2 (54,01%). Sedangkan data dengan kategori "lebih dari 2" paling banyak berada pada *cluster* C3 (54,64). Meskipun pada *cluster* C3 secara jumlah data kategori "lebih dari 2" lebih sedikit dibandingkan dengan kategori "2", tetapi secara persentase sebaran data, kategori "lebih dari 2" lebih tinggi dibandingkan dengan kategori "2". Persentase sebaran data berdasarkan atribut jumlah motor disajikan pada Tabel 11.

Tabel 11. Persentase sebaran data berdasarkan atribut jumlah motor

Kategori	Cluster C1	Cluster C2	Cluster C3
0	12 (40%)	13 (43.33%)	5 (16.67%)
1	**160 (58.18%)**	76 (27.64%)	38 (13.82%)
2	34 (10.09%)	182 (**54.01%**)	119 (35.31%)
lebih dari 2	18 (18.56%)	26 (26.80%)	**53 (54.64%)**

Berdasarkan atribut biaya pulsa dan internet, terlihat bahwa *cluster* C1 didominasi oleh data dengan kategori "0 – 50.000". Data dengan kategori "50.001 – 100.000" merupakan data terbanyak pada *cluster* C2. Sedangkan pada *cluster* C3, data dengan kategori biaya pulsa dan

internet "di atas 200.000" menjadi data yang paling banyak. Distribusi data di dalam *cluster* berdasarkan atribut biaya pulsa dan internet selengkapnya dapat dilihat pada Gambar 8.



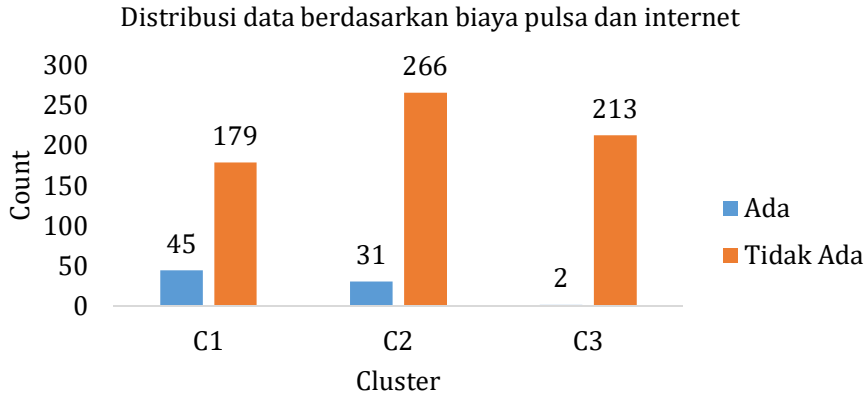
Gambar 8. Distribusi data di dalam *cluster* berdasarkan atribut biaya pulsa dan internet

Berdasarkan persentase sebaran data setiap kategori pada atribut biaya pulsa dan internet, kategori "0 – 50.0000" sebagian besar berada pada *cluster* C1 (65,75%). Sebaran data dengan kategori "50.001 – 100.000" dan "100.001 – 200.000" paling banyak berada pada *cluster* C2, yaitu sebanyak 76,84% dan 49,02%. Sedangkan data dengan kategori "di atas 200.000" paling banyak berada pada cluster C3 yang mencapai 83,33%. Persentase sebaran data berdasarkan atribut biaya pulsa dan internet selengkapnya disajikan pada Tabel 12.

Tabel 12. Persentase sebaran data berdasarkan atribut biaya pulsa dan internet

Kategori	Cluster C1	Cluster C2	Cluster C3
0 - 50.000	**144 (65.75%)**	54 (24.66%)	21 (9.59%)
50.001 - 100.000	30 (15.79%)	**146 (76.84%)**	14 (7.37%)
100.001 - 200.000	43 (28.10%)	**75 (49.02%)**	35 (22.88%)
di atas 200.000	7 (4.02%)	22 (12.65%)	**145 (83.33%)**

Berdasarkan atribut jaminan pendidikan, semua cluster sebagian besar terisi oleh data dengan kategori jaminan pendidikan "Tidak ada". Hal ini terjadi karena distribusi data setiap kategori pada atribut jaminan pendidikan sangat tidak seimbang. Distribusi data di dalam *cluster* berdasarkan atribut jaminan pendidikan dapat dilihat pada Gambar 9.



Gambar 9. Distribusi data di dalam *cluster* berdasarkan atribut jaminan pendidikan

Jika dilihat dari persentase sebaran data untuk setiap kategori, sebaran data berdasarkan atribut jaminan pendidikan dengan kategori "Ada" paling banyak berada pada *cluster* C1 (57,69%) dan sangat sedikit sekali yang berada pada *cluster* C3 (2,57%). Sedangkan kategori "Tidak Ada" hampir merata di semua *cluster*. Persentase sebaran data berdasarkan atribut jaminan pendidikan disajikan pada Tabel 13.

Tabel 13. Persentase sebaran data berdasarkan atribut biaya pulsa dan internet

Kategori	Cluster C1	Cluster C2	Cluster C3
Ada	**45 (57.69%)**	31 (39.74%)	2 (2.57%)
Tidak Ada	179 (27.20%)	**266 (40.43%)**	213 (32.37%)

4. SIMPULAN

Analisis *cluster* untuk pengelompokan mahasiswa berdasarkan latar belakang ekonomi pada data kategorik dapat dilakukan dengan menerapkan metode K-Modes. Dari proses clustering, penentuan jumlah *cluster* dilakukan menggunakan metode *Elbow* dan diperoleh 3 *cluster*. Setiap *cluster* memiliki karakteristik yang berbeda-beda yang ditunjukan oleh kategori-kategori pada setiap atribut yang membentuk *cluster-cluster* tersebut. *Cluster* pertama sebagian besar diisi oleh mahasiswa dengan latar belakang ekonomi yang relatif rendah, cluster kedua diisi oleh mahasiswa-mahasiswa dengan latar belakang ekonomi sedang, dan *cluster* ketiga diisi oleh mahasiswa-mahasiswa dengan latar belakang relatif tinggi.

UCAPAN TERIMA KASIH

Terima kasih penulis haturkan kepada Lembaga penelitian dan Pengabdian kepada Masyarakat UIN Sunan Kalijaga Yogyakarta atas segala dukungan fasilitas yang diberikan. Penelitian ini didasarkan pada penelitian BLU Tahun Anggaran 2023.

REFERENSI

[1] S. N. Mohulaingo, R. Hafid, A. Bahsoan, R. Ilato, and M. Mahmud, "Pengaruh Status Sosial Ekonomi Orang Tua Terhadap Minat Berwirausaha Alumni," *Journal of Economic and Business Education*, vol. 1, no. 1, pp. 1–23, 2023.

[2] J. Taluke, L. Lesawengen, and E. A. A. Suwu, "Pengaruh Status Sosial Ekonomi Orang Tua Terhadap Tingkat Keberhasilan Mahasiswa Di Desa Buo Kecamatan Loloda Kabupaten Halmahera Barat," *Jurnal Holistik*, vol. 14, no. 2, pp. 1–16, 2021, [Online]. Available: https://ejournal.unsrat.ac.id/index.php/holistik/article/view/33777

[3] S. Wakit, "Pengaruh Antara Faktor- Faktor Kesulitan Belajar dan Latar Belakang Sosial Ekonomi Orang Tua Terhadap Prestasi Belajar Mahasiswa Jurusan Tarbiyah Program Studi Manajemen Pendidikan Islam Semester 1 Staida Blokagung -Banyuwangi Tahun Akademik 2010 /2011," *Jurnal Penelitian Iptek*, vol. 2, pp. 14–29, 2017.

[4] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[5] D. Prasetyawan and R. Gatra, "Model Convolutional Neural Network untuk Mengukur Kepuasan Pelanggan Berdasarkan Ekspresi Wajah," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 3, Dec. 2022, doi: 10.28932/jutisi.v8i3.5493.

[6] A. Singh, N. Thakur, and A. Sharma, "A Review of Supervised Machine Learning Algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 1310–1315.

[7] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An Unsupervised Machine Learning Algorithms: Comprehensive Review," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 911–921, 2023, doi: 10.12785/ijcds/130172.

[8] A. Hammoudeh, "A Concise Introduction to Reinforcement Learning," vol. 02, Nov. 2018, doi: 10.13140/RG.2.2.31027.53285.

[9] W. Qiang and Z. Zhongli, "Reinforcement Learning Model, Algorithms and Its application," in *2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)*, 2011, pp. 1143–1146. doi: 10.1109/MEC.2011.6025669.

[10] A. Novoselsky and E. Kagan, *An introduction to cluster analysis*. 2021. doi: 10.13140/RG.2.2.25993.57448/1.

[11] T. Velmurugan, "A State of Art Analysis of Telecommunication Data by k-Means and k-Medoids Clustering Algorithms," *Journal of Computer and Communications*, vol. 06, no. 01, pp. 190–202, 2018, doi: 10.4236/jcc.2018.61019.

[12] Y. Dai, G. Yuan, Z. Yang, and B. Wang, "K-Modes Clustering Algorithm Based on Weighted Overlap Distance and Its Application in Intrusion Detection," *Sci Program*, vol. 2021, p. 9972589, 2021, doi: 10.1155/2021/9972589.

[13] K. Mcilhany and S. Wiggins, "High Dimensional Cluster Analysis Using Path Lengths," *Journal of Data Analysis and Information Processing*, vol. 06, no. 03, pp. 93–125, 2018, doi: 10.4236/jdaip.2018.63007.

[14] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Min Knowl Discov*, vol. 2, no. 3, pp. 283–304, 1998, doi: 10.1023/A:1009769707641.

[15] K. Lakshmi, N. Karthikeyani Visalakshi, S. Shanthi, and S. Parvathavarthini, "CLUSTERING CATEGORICAL DATA USING k-MODES BASED ON CUCKOO SEARCH OPTIMIZATION ALGORITHM," *ICTACT Journal on Soft Computing*, vol. 8, no. 1, pp. 1561–1566, Oct. 2017, doi: 10.21917/ijsc.2017.0218.

[16] M. Á. Carreira-Perpiñán and W. Wang, "The K-modes algorithm for clustering," *ArXiv*, vol. abs/1304.6478, 2013, [Online]. Available: https://api.semanticscholar.org/CorpusID:8655077

[17] F. Indriani and I. Budiman, "K-Modes Clustering untuk Mengetahui Jenis Masakan Daerah yang Populer pada Website Resep Online (Studi Kasus: Masakan Banjar di cookpad.com)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 4, no. 4, p. 290, Dec. 2017, doi: 10.25126/jtiik.201744548.

[18] H. Malikhatin, A. Rusgiyono, and D. A. I. Maruddani, "Penerapan k-Modes Clustering dengan Validasi Dunn Index pada Pengelompokan Karaktereistik Calon TKI Menggunakan R-GUI," vol. 10, no. 3, pp. 359–366, 2021, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/

[19] N. P. M. N. Dewi and I. B. G. Dwidsamara, "Implementation of K-Modes Algorithm for Clustering of Stress Causes in University Students," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 9, no. 3, pp. 419–427, 2021.

[20] D. Desyanti, Y. Yusrizal, and F. Sari, "Implementasi Algoritma K-Modes Untuk Mengukur Tingkat Kepuasan Mahasiswa Terhadap Pembelajaran Daring," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 4, pp. 719–727, Mar. 2022, doi: 10.47065/bits.v3i4.1401.

[21] E. K. Nduru and E. Buulolo, "Implementasi Algoritma K-Modes untuk Menentukan Strategi Marketing STIMIK Budi Darma," *KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer)*, vol. 2, no. 1, 2018, [Online]. Available: http://ejurnal.stmik-budidarma.ac.id/index.php/komik

[22] D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix : Jurnal Manajemen Teknologi dan Informatika*, vol. 9, no. 3, pp. 102–109, Nov. 2019, doi: 10.31940/matrix.v9i3.1662.

[23] A. Aprilliant, "The k-modes as Clustering Algorithm for Categorical Data Type," Medium.com. Accessed: Nov. 20, 2023. [Online]. Available: https://medium.com/geekculture/the-k-modes-as-clustering-algorithm-for-categorical-data-type-bcde8f95efd7