 ISSN: 2684-8473

Kajian Penerapan Jarak Euclidean, Manhattan, Minkowski, dan Chebyshev pada Algoritma Clustering K-Prototype

Rani Nooraeni [1,*], Ghita Nurfalah [2]

Politeknik Statistika STIS, Jakarta, Indonesia
raninoor@stis.ac.id [1], 221810318@stis.ac.id

INFORMASI ARTIKEL

Histori Artikel
Diterima : 07 Juni 2022
Direvisi : 15 Juli 2022
Diterbitkan :19 Agustus 2022

Kata Kunci:
Clustering
KP
FKP
GAFKP
Pengukuran jarak

ABSTRAK

Clustering merupakan teknik data mining yang bertujuan mengelompokkan data yang memiliki kemiripan kedalam satu klaster, semakin tinggi tingkat kemiripan dalam satu klaster semakin baik hasil clustering yang dihasilkan. Kemiripan data tersebut diukur menggunakan fungsi jarak, sehingga memilih fungsi jarak yang tepat sangatlah penting dalam clustering. K-Prototype (KP) adalah algoritma clustering untuk data campuran yang telah banyak digunkan, pengembangan algoritma lainnya dari K-Prototype yang terkenal adalah Fuzzy K-Prototype (FKP) dan Genetic Algorithm K-Prototype (GAFKP). Namun ketiga algoritma tersebut hanya menggunakan jarak Euclidean dalam mengukur kesamaan datanya. Oleh karena itu, dalam penelitian ini dilakukan penerapan jarak Euclidean, Manhattan, Minkowski, dan Chebyshev pada ketiga algoritma tersebut untuk memperoleh kombinasi jarak dan algoritma yang memberikan hasil clustering yang lebih baik. Hasil penelitian menunjukkan bahwa diantara seluruh kombinasi jarak dan algoritma clustering, algoritma Fuzzy K-Prototype dengan jarak Euclidean memberikan hasil yang lebih baik berdasarkan metode evaluasi akurasi dan indeks CV.

I. Pendahuluan

Data mining adalah sebuah teknologi yang memadukan metode analisis data tradisional dengan algoritma canggih untuk memproses data dalam jumlah besar [1]. Dalam *data mining*, *clustering* adalah salah satu teknik paling penting [2]. *Clustering* termasuk dalam algoritma *unsupervised learning* yang bertujuan mempartisi suatu kumpulan data ke dalam kelompok homogen yang disebut klaster [3]. *Clustering* dilakukan dengan mengelompokkan data yang memiliki kemiripan antara satu data dengan data lainnya ke dalam klaster sehingga data dalam satu klaster tersebut memiliki tingkat kemiripan yang maksimal dan data antar klaster memiliki kemiripan yang minimal [4]. Tiga sub masalah yang ditangani oleh *clustering* adalah (i) mendefinisikan pengukuran jarak untuk menilai kesamaan antara elemen data yang berbeda, (ii) menerapkan algoritma yang efisien untuk menemukan klaster dari elemen data yang paling mirip dengan cara *unsupervised*, dan (iii) mendapatkan deskripsi yang dapat mencirikan karakteristik elemen-elemen dari sebuah klaster [3]. Pada umumnya, algoritma clustering yang dibangun hanya dapat digunakan untuk mengelompokkan satu jenis data, seperti data numerik saja, atau data kategorik saja. Namun, dikehidupan nyata data banyak tersedia dalam bentuk campuran numerik dan kategorik [2]. Contoh data yang biasanya bertipe campuran numerik dan kategorik adalah data hasil sensus atau survei [5]. Banyaknya data campuran di kehidupan nyata dikarenakan tidak semua pertanyaan atau permasalahan bisa dijawab dan diselesaikan hanya dengan nilai berskala ukur saja [5].

K-Prototype adalah algoritma *clustering* yang sangat penting untuk melakukan pengelompokkan data campuran numerik dan kategorik [2]. Algoritma ini diusulkan oleh Huang [6] dengan cara mengintegrasikan algoritma *K-Means* dan algoritma *K-Modes*. Pada penelitian selanjutnya, Ji, dkk. [2] menemukan kelemahan dari algoritma *K-Prototype*, sehingga pada penelitiannya dilakukan pengintegrasian *mean* dan *fuzzy centroid* untuk mewakili *prototype* sebuah klaster, serta menggunakan ukuran ketidaksamaan baru untuk mengevaluasi ketidaksamaan antara objek data dan *prototype*. Algoritma yang diusulkan oleh Ji, dkk. kemudian disebut *Fuzzy K-Prototype*. Penelitian berikutnya menemukan bahwa proses inisialisasi centroid yang dilakukan algoritma *Fuzzy K-Prototype* menyebabkan algoritma terjebak dalam solusi *local optimum*. Kelemahan ini kemudian diatasi oleh Arsa [8] dalam penelitiannya yang mengombinasikan algoritma genetika dengan algoritma *Fuzzy*

K-Prototype. Algoritma yang diusulkan oleh Arsa kemudian dinamakan dengan *Genetic Algorithm Fuzzy K-Prototype*.

Dalam proses *clustering*, penentuan derajat kesamaan atau ketidaksamaan data memiliki peran sangat penting, hal ini dilakukan untuk memahami bagaimana data dikatakan saling berhubungan, serupa, dan tidak serupa, sehingga perlu dilakukan analisis komparatif terhadap beberapa metode yang ada [4]. Ketiga algoritma *clustering K-Prototype, Fuzzy K-Prototype*, dan *Genetic Algorithm Fuzzy K-Prototype* sebagai algoritma *clustering* yang dapat menangani data campuran hanya menerapkan jarak Euclidean dalam mengukur kesamaan antara data dengan *centroid*-nya. Walaupun secara umum jarak Euclidean digunakan dalam algoritma *clustering*, namun pilihan jarak lain dalam kasus tertentu dapat dipertimbangkan bergantung pada jumlah data dan kompleksitasnya [9].

Terdapat banyak jenis pengukuran jarak yang dapat digunakan untuk mengukur kesamaan antar data dengan *centroid*-nya. Pilihan pengukuran jarak yang tepat tergantung pada aplikasi dan tipe data yang digunakan [1] [10]. Sementara itu P. Grabusts [9] dalam penelitiannya menyebutkan bahwa jarak Euclidean, jarak Manhattan, jarak Minkowski, dan Jarak Chebyshev merupakan beberapa jarak pengukuran yang dapat digunakan pada algoritma *fuzzy clustering*. Oleh karena itu, pada penelitian ini peneliti melakukan penerapan jarak Euclidean, Manhattan, Minkowski, dan Chebyshev pada ketiga algoritma *K-Prototype, Fuzzy K-Prototype* dan *Genetic Algorithm Fuzzy K-Prototype*. Selanjutnya, peneliti mengkaji hasil dari penerapan keempat jenis jarak pada algoritma *clustering* tersebut dan menggunakan kombinasi jarak dan algorima terbaik yang diperoleh untuk mengelompokkan desa berdasarkan indikator-indikator di dalam Indeks Desa Membangun (IDM) pada data Potensi Desa (Podes) Provinsi Papua tahun 2020.

IDM adalah indeks komposit yang dibentuk dari tiga jenis indeks, yaitu indeks ketahanan sosial, indeks ketahanan ekonomi, dan indeks ketahanan ekologi. Terdapat lima status kemajuan dan kemandirian desa yang ditetapkan dalam IDM, yaitu desa mandiri, desa maju, desa berkembang, desa tertinggal, dan desa sangat tertinggal. Podes adalah data kewilayahan atau spasial yang dimiliki oleh Badan Pusat Statistik (BPS) dengan menekankan pada potensi dari suatu wilayah. Wilayah pendataan Podes mencakup seluruh wilayah administrasi pemerintahan setingkat desa. Podes merupakan dataset yang memiliki ratusan atribut dengan unit observasinya adalah desa sebagai level wilayah terendah [5]. Struktur data Podes yang terdiri dari variabel campuran numerik dan kategorik sangat cocok digunakan pada algoritma *K-Prototype, Fuzzy K-Prototype,* dan *Genetic Algorithm K-Prototype*. Provinsi Papua Barat merupakan wilayah yang memiliki desa dengan status sangat tertinggal kedua paling banyak setelah Provinsi Papua berdasarkan Indeks Desa Membangun yang dirilis oleh Kementerian Desa Pembangunan Daerah Tertinggal dan Transmigrasi Republik Indonesia pada tahun 2020. Pada penelitian ini dilakukan pengelompokkan desa di wilayah Provinsi Papua Barat berdasarkan indikator IDM pada data Podes Provinsi Papua Barat tahun 2020 untuk melihat karakteristik dan potensi yang dimiliki oleh desa di Provinsi Papua Barat tersebut.

II. Material dan Metode

A. Pengumpulan Data

Dalam penelitian ini digunakan dua jenis data, yaitu data uji coba dan data studi kasus. Data uji coba yang digunakan pada penelitian ini merupakan data berlabel yang terdiri dari dua jenis, yaitu data bangkitan dan data *real world*. Data bangkitan terdiri dari 2 jenis data yang dibangkitkan menggunakan aplikasi RStudio.Data bangkitan 1 adalah data dengan jumlah observasi 500 dan jumlah variabel 12, yaitu 8 untuk variabel numerik dan 4 untuk variabel kategorik. Variabel numerik dibangkitkan dengan distribusi normal dan pada setiap variabelnya dibangkitkan dengan beberapa rata-rata yang berbeda. Sedangkan variabel kategorik dibangkitkan dengan distribusi binomial. Variabel yang dibangkitkan dengan distribusi binomial memiliki jumlah kategori yang berbeda dan probabilitas yang beragam pada setiap variabelnya. Variabel terakhir dari data ini merupakan kelas yang membagi data kedalam 2 kelompok. Data bangkitan 2 adalah data dengan jumlah observasi 50 dan jumlah variabel 5, yaitu 3 untuk variabel numerik dan 2 untuk variabel kategorik. Data ini diperoleh dengan cara mengambil sampel sebanyak 50 dan memilih variabel sebanyak 5 dari data bangkitan 1.

Data *real world* yang digunakan diperoleh dari website *UCI Machine Learning Repository* yang terdiri dari 3 jenis, yaitu data *Zoo, data Credit Approval*, dan data *Accute Inflammations*. Data *Zoo* terdiri 17 variabel, yaitu 1 numerik dan 16 kategorik, dengan jumlah observasi 101. Variabel kategorik terakhir merupakan kelas yang membagi data ke dalam 7 kelompok. Data *Credit Approval* terdiri dari 609 observasi dan 16 variabel. Variabel numerik berjumlah 6 dan variabel kategorik berjumlah 10. Variabel kategorik terakhir merupakan kelas yang membagi data menjadi 2 kelompok. Data *Accute Inflammations* memiliki jumlah observasi 120 dengan jumlah variabel 8, yaitu 1 untuk variabel numerik dan 7 untuk variabel kategorik. Dua variabel kategorik terakhir merupakan atribut kelas yang keduanya sama-sama membagi data ke dalam 2 kelas, namun yang atribut kelas yang digunakan dalam penelitian ini adalah atribut kelas pertama atau variabel ke-9.

Data studi kasus yang digunakan pada penelitian ini adalah data Podes Provinsi Papua Barat tahun 2020 yang terdiri dari 1985 observasi dan 24 variabel. Variabel numerik berjumlah 12 dan variabel kategorik berjumlah 12. Data dikelompokkan ke dalam 5 kelas yaitu desa sangat tertinggal, desa tertinggal, desa

berkembang, desa mandiri, dan desa maju. Pemilihan jumlah kelas dan variabel didasarkan pada status desa dan komponen indeks desa membangun yang disesuaikan dengan data yang tersedia pada data potensi desa.

B. Metode Analisis

1. K-Prototype

K-Prototype bertujuan mengelompokkan data kedalam k klaster dengan meminimalkan fungsi objektif berikut ini [2]:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} d(x_i, Q_l) \dots (1)$$

Q_l merepresentasikan vektor atau *prototype* dari sebuah klaster, u_{il} merupakan elemen dari matriks partisi $U_{n\times k}$, dan $d(x_i, Q_l)$ adalah ukuran ketidaksamaan yang didefinisikan dengan:

$$d(x_i, Q_l) = \sum_{j=1}^{p} \left(x_{ij}^r - q_{lj}^r\right)^2 + \mu_l \sum_{j=p+1}^{m} \delta\left(x_{ij}^c, q_{lj}^c\right) \dots (2)$$

$\delta(p, q) = 0$ untuk $p = q$, dan $\delta(p, q) = 1$ untuk $p \neq q$. x_{lj}^r adalah *prototype* dari variabel numerik ke-j didalam klaster l dan x_{lj}^c adalah *prototype* dari variabel kategorik ke-j didalam klaster l. μ_l adalah bobot untuk variabel kategorik di dalam klaster l.

2. Fuzzy K-Prototype

Berikut merupakan fungsi objektif dari algoritma *Fuzzy K-Prototype*:

$$E(U, Q) = \sum_{j=i}^{k} \sum_{i=1}^{n} u_{ij}^{\alpha} d\left(x_i, Q_j\right) \dots (3)$$

Dengan $U = \left(u_{ij}\right)_{n\times k}$ adalah matrix partisi *fuzzy*, yang memenuhi $0 \leq u_{ij} \leq 1$ dan $\sum_{j=1}^{k} \quad u_{ij} = 1$. $Q_j = [q_{j1}, q_{j2}, \dots, q_{jp}, \tilde{v}_{jp+1}, \dots, \tilde{v}_{jp+m}]$ adalah representasi vektor atau *prototype* untuk klaster j; α $(1 < \alpha < \infty)$ adalah koefisien *fuzziness*; dan $d(x_i, Q_j)$ adalah pengukuran ketidaksamaan antara objek data x_i dan *prototype* Q_j, yang didefinisikan:

$$d\left(x_i, Q_j\right) = \sum_{l=1}^{p} \left(w_l\left(x_{il}^r - q_{jl}^r\right)\right)^2 + \sum_{l=p+1}^{m} \varphi\left(x_{il}^c, \tilde{v}_{jl}^c\right)^2) \dots (4)$$

q_{jl}^r adalah *centroid* untuk atribut numerik, $\tilde{v}_{jl}^c$ adalah *fuzzy centroid* untuk atribut kategorik, dan w_l merupakan signifikansi atribut numerik.

3. Genetic Algorithm Fuzzy K-Prototype (GA-FKP)

Proses *Genetic Algorithm Fuzzy K-Prototype* dibagi ke dalam dua tahap, yaitu tahap pertama untuk menjalankan algoritma genetika, dan tahap kedua untuk menjalankan algoritma *Fuzzy K-Prototype*.

Tahap Pertama (Algoritma Genetika)

1. Input parameter, parameter yang perlu diinput adalah data, jumlah klaster (k), nilai koefisien *fuzziness* (α), jumlah populasi, maksimum iterasi, dan *mutation rate*.
2. Inisial populasi, inisialisasi populasi dilakukan dengan membangkitkan kromosom dengan algoritma yang dikenalkan oleh Zhao, Tsujimura, dan Gen (1996) [8].
3. Evaluasi nilai *fitness*, pada tahap ini dilakukan evaluasi nilai *fitness* setiap kromosom untuk mengukur seberapa baik solusi yang diberikan. Semakin tinggi nilai *fitness*, maka solusi yang dimiliki kromosom semakin baik.
4. Membuat populasi baru, pembuatan populasi baru dilakukan melalui tiga tahap, yaitu seleksi, *crossover*, dan mutasi.
5. Menjalankan algoritma dengan populasi baru.

6. Lakukan pengulangan dari tahap ke-3 hingga mencapai iterasi maksimal.

Tahap Kedua (Fuzzy K-Prototype)

1. Dari tahap pertama yang telah dilakukan akan diperoleh matriks partisi *fuzzy*. Matriks dengan nilai *fitness* terbaik tersebut kemudian digunakan pada algoritma *Fuzzy K-Prototype*.
2. Selanjutnya dilakukan penghitungan nilai *centroid* untuk data numerik dan kategorik serta penghitungan jarak setiap objek data pada *centroid* tersebut.
3. Diperoleh hasil berupa klaster.

C. Ukuran Jarak

Jarak Euclidean

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - y_{jk}\right)^2} \ldots (5)$$

d adalah jarak antara i dan j, i sebagai klaster data pusat, j data dalam atribut, k adalah symbol dari setiap data, n adalah banyak data, x_{ik} adalah data dalam klaster pusat hingga k, y_{ik} adalah setiap objek data hingga ke k

Jarak Manhattan

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - y_{jk}| \ldots (6)$$

Jarak Minkowski

$$d_{ij} = \left(\sum_{k=1}^{n} \left|x_{ik} - y_{jk}\right|^p\right)^{\frac{1}{p}} \ldots (7)$$

p adalah bilangan bulat positif (*integer*)

Jarak Minkowski adalah generalisasi dari jarak Euclidean, dan Manhattan [7]. Ketika p pada jarak Minkowski bernilai 1 maka akan menjadi jarak Manhattan. Lalu ketika p bernilai 2 maka akan menjadi jarak Euclidean. Sedangkan ketika $p = \infty$ akan menjadi jarak Chebyshev. Ketika $p = 3, 4, 5\ dan\ \infty$ nilai matriks akan lebih kecil [1].

Jarak Chebyshev

$$d_{ij} = |x_{ik} - y_{jk}| \ \ldots (8)$$

D. Metode Evaluasi

1. Akurasi

Akurasi *clustering* (r) didefinisikan dengan:

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \ldots (9)$$

Dari rumus akurasi di atas, a_i merupakan jumlah objek data yang sesuai antara hasil clustering pada kelas ke-i dengan kelas sebenarnya, dan n merupakan jumlah dari seluruh objek data di dalam data set. Semakin tinggi nilai akurasi mengindikasikan semakin baik hasil clustering yang diperoleh, dan nilai maksimal $r = 1$ yang menunjukkan hasil clustering yang sempurna.

2. Indeks Validitas CV

Indeks validitas CV mengombinasikan fungsi *categoriy utility* (CU) dengan varians yang dapat digunakan untuk megevaluasi efektifitas algoritma *clustering* (Hsu & Chen, 2007). CU digunakan untuk mengukur tingkat homogenitas data kategorik, semakin tinggi CU semakin baik cluster yang dihasilkan. Sedangkan varians digunakan untuk mengevaluasi kualitas *clustering* data numerik, semakin kecil nilai varians, semakin baik hasil *clustering*. Semakin tinggi CV, *cluster* yang dihasilkan semakin baik.

$$CU = \sum_{k} \left(\frac{|C_k|}{D} \sum_{i} \sum_{j} \left[P(A_j = V_{ij}|C_k \right]^2 - P\left(A_j = V_{ij}\right)^2 \right) \dots (10)$$

D adalah ukuran dari *dataset*. $P(A_j = V_{ij}|C_k)$ adalah peluang bersyarat bahwa atribut i memiliki nilai V_{ij} pada klaster C_k dan $P(A_j = V_{ij})$ adalah peluang atribut i memiliki nilai V_{ij} pada seluruh data.

$$\sigma^2 = \sum_{k} \frac{1}{|C_k|} \sum_{i} \sum_{j} \left(V_{i,j}^k - V_{i,avg}^k\right) \dots (11)$$

$V_{i,avg}^k$ adalah rata-rata atribut i dalam klaster k. dan $V_{i,j}^k$ adalah data ke-j untuk atribut i dalam klaster k.

$$CV = \frac{CU}{1 + Variance} \dots (12)$$

3. Waktu Komputasi

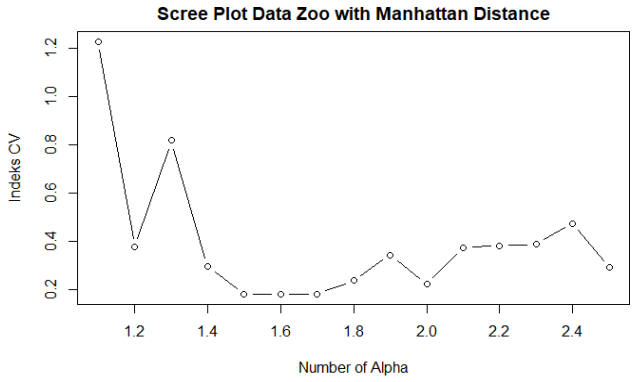
Waktu komputasi diperoleh dengan menggunakan fungsi 'sys.time()' atau waktu sistem pada RStudio, yaitu selisih dari waktu mulai menjalankan algoritma untuk mengklasterkan data dengan waktu selesai menjalankan algorima hingga hasil klaster diperoleh. Semakin kecil waktu komputasi yang diperoleh menunjukkan bahwa algoritma clustering semakin cepat dalam memberikan hasil pengklasteran data. Sedangkan semakin besar waktu komputasi yang diperoleh menunjukkan bahwa algoritma clustering semakin lambat dalam memberikan hasil pengklasteran data.

III. Hasil dan Pembahasan

A. Penentuan Parameter Algoritma

Pada algoritma *K-Prototype*, dibutuhkan parameter k atau jumlah kelas dari data yang akan diklaster. Dalam penelitian ini, karena data uji coba yang digunakan merupakan data berlabel yang telah memiliki kelas, maka nilai k disesuaikan dengan jumlah kelas yang telah tersedia pada masing-masing data. Pada data *Zoo*, terdapat variabel kategorik pada data yang berisi kelas dan membagi data kedalam tujuh kelompok, sehingga nilai k untuk data Zoo adalah $k = 7$. Pada data *credit approval* dan *acute inflammations*, variabel kategorik terakhirnya membagi data kedalam dua kelas, sehingga nilai k untuk data *credit approval* dan *acute inflammations* adalah $k = 2$. Sedangkan untuk data bangkitan dengan jumlah observasi 500 dan 50 juga dibagi kedalam dua kelas berdasarkan pada variabel kategorik terakhirnya, sehingga nilai k untuk data bangkitan dengan jumlah observasi 500 dan 50 adalah $k = 2$.

Pada algoritma *Fuzzy K-Prototype*, parameter yang dibutuhkan untuk menjalankan algoritma adalah jumlah klaster k, iterasi maksimal, koefisien *fuzziness* (α), interval T, dan nilai ε. Untuk parameter jumlah klaster (k) pada algoritma *Fuzzy K-Prototype* digunakan nilai yang sama dengan jumlah klaster pada algoritma *K-Prototype*. Sedangkan untuk parameter lainnya ditetapkan iterasi maksimal atau $Ite = 100$, interval $T = 10$, dan nilai $\varepsilon = 1 \times 10^{-5}$. Sementara itu, untuk nilai koefisien *fuzziness* (α) dicari dengan menggunakan *scree plot* dari indeks CV. Berikut merupakan proses penentuan nilai α untuk setiap data.



Gambar 1. *Scree plot* α data *zoo* untuk jarak Manhattan

Gambar 1 memperlihatkan *scree plot* data *Zoo* untuk jarak Manhattan. Terlihat jelas pada *scree plot* bahwa nilai indeks CV tertinggi berada pada nilai $\alpha = 1.1$, dengan nilai indeks CV nya 1.2271576. Oleh karena itu,

pada jarak Manhattan ini dipilih nilai $\alpha = 1.1$. Oleh karena dalam penelitian ini digunakan 3 jenis algoritma *clustering* dan 4 jenis pengukuran jarak, maka terdapat 12 kombinasi jarak dan algoritma tersebut. Pada seluruh kombinasi tersebut dilakukan penentuan nilai parameter alpha dengan cara yang sama, yaitu menggunakan *scree plot* nilai alpha dan indeks CV.

Selanjutnya, parameter pada *Genetic Algorithm Fuzzy K-Prototype* adalah jumlah klaster k, koefisien *fuzziness* (α), jumlah populasi, maksimum iterasi, dan *mutation rate*. Selain jumlah populasi, parameter pada *Genetic Algorithm Fuzzy K-Prototype* merupakan parameter yang sama seperti pada *Fuzzy K-Prototype*. Oleh karena itu, nilai parameter yang digunakan pada *Genetic Algorithm Fuzzy K-Prototype* menggunakan nilai yang sama seperti yang digunakan pada *Fuzzy K-Prototype*. Sedangkan untuk jumlah populasi, digunakan jumlah populasi sebanyak 20.

B. Hasil Clustering Data Uji Coba

Tabel di bawah ini merupakan hasil *clustering* dari seluruh jarak dan algoritma untuk semua data uji coba yang telah dievaluasi menggunakan akurasi. Nilai akurasi diperoleh dengan mencocokkan banyaknya data hasil *clustering* yang memiliki kelas yang bersesuaian dengan kelas sebenarnya. Semakin tinggi nilai akurasi menunjukkan hasil *clustering* yang semakin baik, karena hal ini menunjukkan semakin banyak pula objek data yang tepat diklasterkan sesuai dengan kelas sebenarnya.

Tabel 1. Evaluasi Akurasi dari Hasil Clustering pada Data Uji Coba

Algoritma	Zoo	Inflammations	Credit	Bangkitan 1	Bangkitan 2
KPEuc	0.6832	0.7417	0.8033	0.9980	0.5833
KPMnh	0.8119	0.6604	0.8468	0.7320	1.0000
KPMnk	0.5347	0.3546	0.8018	0.7320	0.9400
KPCby	0.7228	0.7417	0.6517	0.5640	0.8000
FKPEuc	0.9505	0.4167	0.8514	1.0000	1.0000
FKPMnh	0.8515	0.5000	0.7387	1.0000	1.0000
FKPMnk	0.8713	0.7583	0.8108	0.7280	0.8200
FKPCby	0.7129	0.4167	0.9474	0.7280	0.8000
GAFKPEuc	0.9208	0.7583	0.8153	1.0000	0.9600
GAFKPMnh	0.9010	0.6667	0.7387	1.0000	1.0000
GAFKPMnk	0.9010	0.7583	0.8108	0.7280	0.8200
GAFKPCby	0.6040	0.4167	0.7988	0.7280	0.7000

Pada tabel 1 terlihat bahwa data *Zoo* memiliki nilai akurasi tertinggi ketika diklasterkan oleh algoritma FKP dengan jarak Euclidean dengan nilai akurasi 0.9505. Data *Accute Inflammations* memiliki akurasi tertinggi sebesar 0.7583 ketika diklasterkan oleh algoritma GAFKP dengan jarak Euclidean. Data *Credit Approval* memiliki nilai akurasi tertinggi 0.9474 ketika diklasterkan oleh algoritma FKP dengan jarak Chebyshev. Data bangkitan 1 memperoleh akurasi sempurna, yaitu bernilai 1 ketika diklasterkan oleh algoritma FKP dengan jarak Euclidean, FKP dengan jarak Manhattan, GAFKP dengan jarak Euclidean, dan GAFKP dengan jarak Manhattan. Data bangkitan 2 juga memperoleh nilai akurasi sempurna sebesar 1 ketika diklasterkan oleh algoritma KP dengan jarak Manhattan, FKP dengan jarak Euclidean, FKP dengan jarak Manhattan, dan GAFKP dengan jarak Manhattan.

Berdasarkan tabel I tersebut, algoritma FKP dengan jarak Euclidean berhasil memberikan tiga nilai akurasi tertinggi dari lima data uji coba yang diklasterkan. Oleh karena itu, algoritma FKP dengan jarak Euclidean menjadi algorima terbaik berdasarkan evaluasi nilai akurasi.

Tabel 2. Evaluasi Indeks CV dari Hasil Clustering pada Data Uji Coba

Algoritma	Zoo	Inflammations	Credit	Bangkitan 1	Bangkitan 2
KPEuc	0.3462	0.0891	9.1209E-09	0.0048	0.0084
KPMnh	0.2700	0.0841	1.0637E-08	0.0019	0.0034
KPMnk	0.2309	0.1631	1.1227E-08	0.0019	0.0061
KPCby	0.2846	0.0891	1.7981E-08	0.0026	0.0030
FKPEuc	0.2625	0.1631	1.1363E-08	0.0048	0.0074
FKPMnh	0.7851	0.1541	4.5700E-09	0.0048	0.0078
FKPMnk	0.2671	0.1631	1.1422E-08	0.0037	0.0049
FKPCby	0.2748	0.1631	1.1362E-08	0.0037	0.0048
GAFKPEuc	0.2625	0.1631	1.1373E-08	0.0048	0.0074
GAFKPMnh	0.2355	0.1539	4.4246E-09	0.0048	0.0078

GAFKPMnk	0.2577	0.1631	1.1422E-08	0.0037	0.0049
GAFKPCby	0.0819	0.1631	1.0734E-08	0.0037	0.0034

Tabel 2 memperlihatkan evaluasi hasil *clustering* menggunakan indeks CV pada seluruh kombinasi jarak dan algoritma untuk kelima data uji coba. Semakin tinggi nilai indeks CV menunjukkan hasil *clustering* yang semakin baik. Pada data *Zoo*, indeks CV tertinggi sebesar 0.7851 diperoleh ketika data diklasterkan oleh algoritma FKP dengan jarak Manhattan. Pada data *Accute Inflammations* indeks CV memiliki nilai yang sama besar yaitu 0.1631 ketika data diklasterkan oleh algoritma KP dengan jarak Minkowski, algoritma FKP dengan jarak Euclidean dan Minkowski, dan algoritma GAFKP dengan jarak Euclidean, Minkowski, dan Chebyshev. Pada data *Credit Approval* indeks CV tertinggi diperoleh ketika data diklasterkan menggunakan algoritma KP dengan jarak Chebyshev dengan nilai indeks CV sebesar 1.7981E-08 atau 0,000000017981. Pada data bangkitan 1, terdapat indeks CV dengan nilai yang sama besar, yaitu 0.0048 pada algoritma KP dengan jarak Euclidean, algoritma FKP dengan jarak Euclidean dan Manhattan, serta algoritma GAFKP dengan jarak Euclidean dan Manhattan. Sedangkan pada data bangkitan 2, indeks CV tertinggi diperoleh ketika data diklasterkan menggunakan algoritma KP dengan jarak Euclidean dengan nilai indeks CV sebesar 0.0084.

Berdasarkan tabel II di atas, algoritma KP dengan jarak Euclidean, FKP dengan jarak Euclidean, FKP dengan jarak Manhattan, dan GAFKP dengan jarak Euclidean memiliki dua nilai indeks CV tertinggi diantara 5 data uji coba. Oleh karena itu, algoritma-algoritma tersebut merupakan algoritma terbaik apabila dilihat dari hasil evaluasi indeks CV

Tabel 3. Evaluasi Waktu Komputasi pada Data Uji Coba

Algoritma	Zoo	Inflammations	Credit	Bangkitan 1	Bangkitan 2
KPEuc	0.7750	0.5118	0.3348	0.5414	0.1097
KPMnh	0.5264	0.4520	0.1616	0.5074	0.0659
KPMnk	0.6014	0.3989	0.1674	0.4800	0.0658
KPCby	1.3604	0.3870	0.1476	0.4024	0.6843
FKPEuc	1.6276	0.1735	0.1546	0.1546	0.0878
FKPMnh	0.4857	0.0977	0.2753	0.1107	0.1052
FKPMnk	1.6775	0.2257	0.6497	0.1556	0.1057
FKPCby	0.2573	0.0808	0.1082	0.1183	0.0499
GAFKPEuc	298.8403	148.3735	166.9331	166.1126	68.9041
GAFKPMnh	209.9315	148.7036	208.3835	99.8351	92.8189
GAFKPMnk	205.8035	152.5520	216.8454	98.4296	91.0376
GAFKPCby	298.6400	154.7139	200.8341	133.0852	100.4136

Tabel 3 memperlihatkan waktu komputasi dalam satuan detik dari seluruh kombinasi jarak dan algoritma *clustering* selama proses mengklasterkan data uji coba. Semakin kecil waktu komputasi menunjukkan bahwa algoritma *clustering* semakin cepat dalam mengklasterkan data. Terlihat bahwa pada data *Zoo, Accute Inflammations, Credit Approval*, dan data bangkitan 2 waktu komputasi tercepat diperoleh ketika data data dikelasterkan oleh algoritma FKP dengan jarak Chebyshev, dengan waktu yang dibutuhkan untuk mengklasterkan masing-masing data adalah 0.2573 detik, 0.0808 detik, 0.1082 detik, dan 0.0499 detik. Sedangkan pada data bangkitan 1 waktu komputasi tercepat diperoleh ketika menggunakan algoritma FKP dengan jarak Manhattan dengan waktu komputasi 0.1107 detik.

Berdasarkan tabel 3 di atas, algoritma FKP dengan jarak Chebyshev menjadi algoritma dengan waktu komputasi tercepat dalam mengklasterkan seluruh data uji coba kecuali data bangkitan 1. Oleh karena itu, algoritma FKP dengan jarak Chebyshev menjadi algoritma terbaik berdasarkan hasil evaluasi waktu komputasi.

Pengaruh jumlah variabel dan observasi

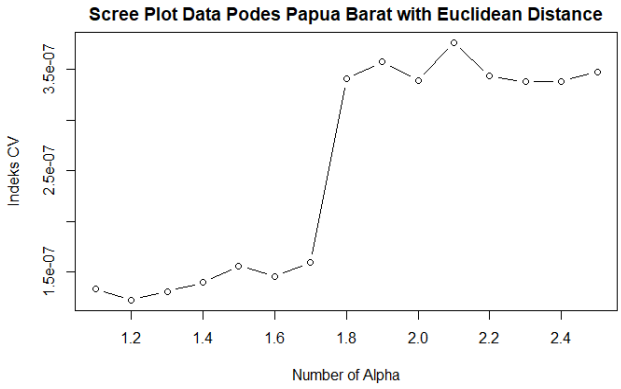
Data bangkitan 1 dan 2 dibuat untuk dapat melihat pengaruh jumlah variabel dan jumlah observasi terhadap hasil *clustering*. Dari hasil evaluasi akurasi, terlihat bahwa data bangkitan 2 dengan jumlah observasi dan variabel yang lebih sedikit dari data bangkitan 1, yaitu 50 observasi dan 5 variabel menunjukkan nilai akurasi yang sedikit lebih tinggi darpada data bangkitan 1 dengan jumlah observasi 500 dan jumlah variabel 12. Pada indeks CV juga terlihat bahwa data bangkitan 2 memiliki nilai indeks CV yang sedikit lebih tinggi disbanding nilai indeks CV pada data bangkitan 1. Selain itu, jumlah observasi dan variabel yang banyak juga berpengaruh pada waktu komputasi yang lebih lama dibandingkan dengan data dengan jumlah observasi dan variabel yang sedikit.

Menentukan kombinasi jarak dan algoritma terbaik

Penentuan kombinasi jarak dan algoritma terbaik dapat diperoleh dengan menggabungkan hasil evaluasi akurasi, indeks CV, dan waktu komputasi. Berdasarkan evaluasi akurasi dan indeks CV, algoritma *Fuzzy K-Prototype* dengan jarak Euclidean lebih unggul dibandingkan dengan algoritma dan jarak lainnya. Sedangkan dari sisi kecepatan waktu komputasi, algoritma *Fuzzy K-Prototype* dengan jarak Chebyshev lebih unggul diantara algoritma dan jarak lainnya. Namun, apabila dilihat secara keseluruhan, maka algoritma *Fuzzy K-Prototype* dengan jarak Euclidean lah yang paling unggul diantara kombinasi jarak dan algoritma lainnya. Oleh karena itu, pada penelitian ini ditetapkan bahwa algoritma *Fuzzy K-Prototype* dengan jarak Euclidean merupakan kombinasi jarak dan algoritma yang memberikan hasil *clustering* terbaik. Selanjutnya, algoritma *Fuzzy K-Prototype* dengan jarak Euclidean inilah yang akan digunakan untuk mengelompokkan desa di Provinsi Papua Barat berdasarkan data Podes Provinsi Papua Barat tahun 2020.

C. Hasil Clustering Data Studi Kasus

Provinsi Papua Barat terdiri dari 13 Kabupaten/Kota, 218 Kecamatan, dan 1987 desa. Namun, dalam data Podes Papua Barat tahun 2020 yang digunakan dalam penelitian ini hanya terdapat 1985 desa. Parameter yang digunakan dalam algoritma *Fuzzy K-Prototype* untuk mengelompokkan desa di Provinsi Papua Barat pada dasarnya sama dengan parameter yang digunakan untuk mengelompokkan data uji coba, hanya saja terdapat perbedaan nilai pada parameter jumlah klaster k dan nilai koefisien *fuzziness* (α). Jumlah klaster yang digunakan untuk mengelompokkan desa berdasarkan data Podes Provisinsi Papua Barat ini adalah $k = 5$. Jumlah klaster ini digunakan berdasarkan jumlah status kemajuan desa pada Indeks Desa Membangun, dimana terdapat lima status desa, yaitu desa mandiri, desa maju, desa berkembang, desa tertinggal, dan desa sangat tertinggal. Sedangkan untuk nilai α ditentukan dengan melihat *scree plot* dari indeks CV berikut.



Gambar 2. *Scree plot* α data Podes Papua Barat

Gambar 2 menunjukkan *scree plot* dari nilai α dan indeks CV untuk data Podes Provinsi Papua Barat tahun 2020. Pada *scree plot* terlihat bahwa indeks CV tertinggi sebesar 3.766485e-07 atau 0.0000003766485s terdapat pada nilai $\alpha = 2.1$. Oleh karena itu, nilai α yang digunakan pada *clustering* data Podes Provinsi Papua Barat tahun 2020 ini adalah $\alpha = 2.1$.

Berikut merupakan hasil *clustering* data Podes Provinsi Papua Barat tahun 2020 oleh algoritma *Fuzzy K-Prototype* dengan jarak Euclidean.

Tabel 4. Jumlah Anggota Klaster pada Data Studi Kasus

Klaster	Jumlah Anggota (desa)	Persentase (%)
Klaster 1	8	0.4030
Klaster 2	539	27.1536
Klaster 3	34	1.7128
Klaster 4	1365	68.7657
Klaster 5	39	1.9647

Tabel 4 memperlihatkan jumlah anggota klaster dari data Podes Papua Barat, yaitu klaster 1 terdiri dari 8 desa, klaster 2 terdiri dari 539 desa atau sekitar 27.15% dari jumlah desa di Papua Barat berada di klaster 2,

lalu klaster 3 terdiri dari 34 desa, klaster 4 dengan jumlah desa terbanyak yaitu 1365 desa atau 68.76%, dan klaster 5 yang terdiri dari 39 desa.

Karakteristik desa hasil *clustering*

Tabel 5. Karakteristik Desa Hasil Clustering Berdasarkan Variabel Numerik

Variabel	Klaster				
	1	2	3	4	5
1	65.25	390.65	64.29	41.71	69.49
2	17.25	2.23	1.62	16.08	5.15
3	1.63	3.43	1.18	0.71	1.36
4	1.50	2.73	1.35	0.84	1.44
5	0.25	4.98	2.06	0.72	1.41
6	0.00	0.30	0.00	0.01	0.00
7	0.13	0.19	0.00	0.02	0.00
8	2.63	18.85	3.91	1.85	3.41
9	0.13	0.12	0.00	0.03	0.05
10	1.00	3.79	1.26	0.78	1.33
11	1.13	1.62	0.56	0.79	0.82
12	0.38	0.28	0.26	0.07	0.18

Tabel 5 memperlihatkan karakteristik desa hasil *clustering* berdasarkan rata-rata dari variabel numerik. Pada variabel pertama yang menunjukkan jumlah keluarga pengguna listrik, terlihat bahwa klaster 2 memiliki jumlah keluarga pengguna listrik tertinggi, disusul oleh klaster 5, kemudian klaster 1, klaster 3, dan klaster 4. Sedangkan pada jumlah keluarga bukan pengguna listrik. Variabel 8 mengenai sarana lembaga keuangan, klaster 2 menyatakan terdapat sekitar 18.85 (18/19) unit. Variabel 10 menyatakan jumlah sarana/prasarana ekonomi, klaster 2 menyatakan terdapat 3 sampai 4 sarana ekonomi, dan klaster lainnya dengan 1 hingga 2 sarana ekonomi.

Tabel 6. Karakteristik Desa Hasil Clustering Berdasarkan Variabel Kategorik

Variabel	Klaster				
	1	2	3	4	5
1	3 (37.5%)	2 (41.18%)	3 (58.82%)	3 (82.63%)	2 (43.58%)
2	5 (100%)	4 (50.83%)	5 (91.17%)	5 (97.21%)	5 (92.30%)
3	2 (75%)	2 (76.62%)	2 (91.17%)	2 (75.09%)	2 (79.48%)
4	1 (75%)	1 (86.08%)	1 (79.41%)	1 (36.33%)	1 (74.35%)
5	7 (50%)	6 (33.02%)	6 (41.17%)	7 (41.46%)	7 (41.02%)
6	4 (62.5%)	4 (48.05%)	4 (58.82%)	6 (51.28%)	4 (38.46%)
7	2 (50%)	1 (65.30%)	1 (44.11%)	1 (71.57%)	1 (56.41%)
8	3 (62.5%)	1 (47.12%)	3 (44.11%)	2 (50.76%)	1 (35.89%)
9	2 (100%)	2 (77.92%)	2 (100%)	2 (97.87%)	2 (92.30%)
10	4 (50%)	2 (43.22%)	3 (50%)	4 (41.53%)	2 (41.02%)
11	4 (100%)	4 (86.27%)	4 (100%)	4 (99.70%)	4 (100%)
12	2 (87.5%)	2 (86.45%)	2 (97.05%)	2 (95.45%)	2 (84.61%)

Tabel 6 memperlihatkan karakteristik desa hasil *clustering* berdasarkan modus dan persentase dari nilai di dalam variabel kategorik. Pada variabel 1 mengenai penerangan di jalan utama desa/kelurahan, hanya klaster 2 dan 5 saja yang menyatakan ada walau hanya sebagian kecil, sedangkan pada klaster 3 82.63% menyatakan tidak ada. Pada variabel 2, yaitu bahan bakar untuk memasak, seluruh klaster menyatakan sebagian besar memasak dengan kayu bakar, kecuali klaster 2 yang menggunakan minyak tanah sebagai bahan bakar memasak. Pada variabel 5 mengenai sumber air minum, klaster 2 dan 3 menggunakan sumur, sedangkan klaster lainnya menggunakan mata air untuk sumber air minum. Variabel 7 mengenai lalu lintas, 50% menyatakan melalui air, sedangkan empat klaster lainnya memilih darat untuk lalu lintas dari/ke desa/kelurahan. Variabel 8 tentang angkutan umum, kelas 2 dan 5 menyatakan ada dengan trayek yang tepat, sedangkan kelas 4 menyatakan ada dengan trayek tidak tetap. Untuk variabel 10, klaster 2 dan 5 menyatakan bahwa terdapat sinyal telepon yang kuat di wilayahnya. Setelah melakukan identifikasi terhadap karakteristik desa berdasarkan variabel numerik dan kategorik, dapat disimpulkan bahwa klaster 2 merupakan desa mandiri, klaster 5 merupajan desa maju, klaster 3 merupakan desa berkembang, klaster 1 merupakan desa tertinggal, dan klaster 4 merupakan desa sangat tertinggal.

IV. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa dari penerapan jarak Euclidean, Manhattan, Minkowski, dan Chebyshev pada algoritma *K-Prototype*, *Fuzzy K-Prototype*, dan *Genetic Algorithm K-Prototype* diperoleh kombinasi jarak dan algoritma terbaik berdasarkan hasil evaluasi akurasi, indeks CV, dan waktu komputasi. Algoritma dan jarak terbaik yang diperoleh yaitu algoritma *Fuzzy K-Prototype* dengan jarak Euclidean. Selanjutnya, hasil *clustering* data Podes Papua Barat tahun 2020 oleh algoritma *Fuzzy K-Prototype* dengan jarak Euclidean menunjukkan Provinsi Papua terdiri dari 8 desa tertinggal atau sekitar 0.40%, 539 desa mandiri atau sekitar 27.15%, 34 desa berkembang atau sekitar 1.71%, 1365 desa sangat tertinggal atau sekitar 68.76%, dan 39 desa maju atau sekitar 1.96%.

Ucapan Terima Kasih

Bagian ini untuk mengucapkan terima kasih kepada pihak-pihak yang telah membantu penerbitan paper ini.

Daftar Pustaka

[1] Thant, Aye & Aye, Soe. (2020). Euclidean, Manhattan and Minkowski Distance Methods For Clustering Algorithms. International Journal of Scientific Research in Science, Engineering and Technology. 553-559. 10.32628/IJSRSET2073118.

[2] Ji, J., Pang, W., Zhou, C., Han, X., & Wang, Z. (2012). A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. Knowl. Based Syst., 30, 129-135.

[3] Ahmad, Amir & Dey, Lipika. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering. 63. 503-527. 10.1016/j.datak.2007.03.016.

[4] Faisal, M & Zamzami, E & Sutarman, (2020). Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance. Journal of Physics: Conference Series. 1566. 012112. 10.1088/1742-6596/1566/1/012112.

[5] Nooraeni, R. (2015). Metode Cluster Menggunakan Kombinasi Algoritma Cluster K-Prototype dan Algoritma Genetika untuk Data Bertipe Campuran.

[6] Huang, Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values.

[7] Khairi, R & Fitri, Sari & Rustam, Zuherman & Pandelaki, Jacub. (2021). Fuzzy C-Means Clustering with Minkowski and Euclidean Distance for Cerebral Infarction Classification. Journal of Physics: Conference Series. 1752. 012033. 10.1088/1742-6596/1752/1/012033.

[8] Arsa, M.I. (2018). Kombinasi Algoritme Genetika dan Fuzzy K-Prototype untuk Pengelompokan Data Campuran.

[9] Grabusts, Peter. (2015). The Choice of Metrics for Clustering Algorithms. Environment. Technology. Resources. Proceedings of the International Scientific and Practical Conference. 2. 70. 10.17770/etr2011vol2.973.

[10] Bora, Dibya & Gupta, Dr. (2014). Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab. 5.

[11] Liu, Hsiang-Chuan & Jeng, Bai-Cheng & Yih, Jeng-Ming & Yu, Yen-Kuei. (2009). Fuzzy C-means algorithm based on standard mahalanobis distances. Proceedings of the 2009 International Symposium on Information Processing (ISIP'09).

[12] Ahmad, A., & Khan, S.S. (2019). Survey of State-of-the-Art Mixed Data Clustering Algorithms. IEEE Access, 7, 31883-31902.

[13] Haryati, A. E., Surono, S., & Suparman, S. (2021). Implementation of Minkowski-Chebyshev Distance in Fuzzy Subtractive Clustering. EKSAKTA: Journal of Sciences and Data Analysis, 2(2), 82–87. https://doi.org/10.20885/EKSAKTA.vol2.iss2.art1.

[14] Hsu, Chung-Chian & Huang, Yan-Ping. (2008). Incremental clustering of mixed data based on distance hierarchy. Expert Systems with Applications. 35. 1177-1185. 10.1016/j.eswa.2007.08.049.

[15] Ji, Jinchao & Zhou, Chunguang & Wang, Zhe & He, Jialiang & Bai, Tian. (2012). A fuzzy k-prototypes algorithm using fuzzy centroid for clustering mixed data. International Journal of Advancements in Computing Technology. 4. 281-290. 10.4156/ijact.vol4.issue7.31.

[16] Nishom, M.. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. Jurnal Informatika: Jurnal Pengembangan IT. 4. 20-24. 10.30591/jpit.v4i1.1253.

[17] Nooraeni, R., Arsa, M.I., & Kusumo Projo, N.W. (2021). Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering. Procedia Computer Science, 179, 677-684.

[18] Santoso, A.B. (2021). Fuzzy K-Prototype Geographically Weighted Clustering yang Dioptimasi Menggunakan Algoritma Genetika untuk Data Campuran (Studi Kasus: Indikator Indeks Pembangunan Desa di Kabupaten Temanggung Tahun 2018).

[19] Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PloS one, 10(12), e0144059. https://doi.org/10.1371/journal.pone.0144059.

[20] Singh, Archana & Yadav, Avantika & Rana, Ajay. (2013). K-means with Three different Distance Metrics. International Journal of Computer Applications. 67. 13-17. 10.5120/11430-6785.

[21] Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. R J., 10, 200.

[22] Widodo, S., Brawijaya, H., & Samudi, S. (2021). Clustering Kanker Serviks Berdasarkan Perbandingan Euclidean dan Manhattan Menggunakan Metode K-Means.