Analisis Algoritma Partitioning Around Medoid untuk Penentuan Klasterisasi

Mira Orisa
Program Studi S1 Teknik Informatika
Fakultas Teknologi Industri Institut
Teknologi Nasional (ITN) Malang
Malang, Indonesia
mir4orisa@gmail.com

Ahmad Faisol
Program Studi S1 Teknik Informatika
Fakultas Teknologi Industri Institut
Teknologi Nasional (ITN) Malang
Malang, Indonesia
mzfais@lecturer.itn.ac.id

***Abstract*— Algoritma *Partitioning Around Medoid* dikenal dengan K-medoids. Algoritma K-Medoids lebih cocok digunakan pada Dataset yang memiliki *outlier*. Karena K-Medoids merupakan perbaikan dari algoritma K-Means pada *clustering* yang kurang baik dalam menangani dataset yang memiliki *outliers*. Algoritma K-Medoids menentukan pusat *cluster* berdasarkan perwakilan objek *cluster* yang disebut dengan medoid. Medoid adalah objek *cluster* yang terletak paling sentral, dengan jumlah jarak minimum ke titik lain. Untuk menutupi kelemahan metode K-Medoids dalam menentukan jumlah k awal secara random digunakan metode *elbow*. Evaluasi kelayakan algoritma K-Medoids dalam pembentukan klasterisasi dilakukan pengukuran terhadap silhouette coefficient dan Davies-Bouldin index.**

Keywords*-*PAM, K-Medoids, elbow,silhouette, davies-Bouldin index

PENDAHULUAN

Teknik klasterisasi yang umum digunakan adalah K-Means. Pada Tahun 2019, Gustientiedina dan kawan-kawan melakukan penelitian klasterisasi data obat-obatan pada RSUD Pekanbaru menggunkaan algoritma K-Means[1]. Dan Annur pada tahun yang sama juga melakukan penelitian untuk mengelompokkan data produk yang terjual menggunakan algoritma K-Means[2]. Dan Mustofa melakukan penelitian untuk mengklasterisasi karakter permainan multiplayer online battle arena menggunakan algoritma K-Means[3]. Pada tahun 2020,Siringoringo dan kawan-kawan melakukan penelitian untuk mengklasterisasi topik berita dengan algoritma K-Means[4]. Masih di tahun 2020, penelitian tentang meningkatkan protokol perutean OLSR menggunakan pengelompokan K-means oleh Y. Hamzaoui dan kawan-kawan[5]. Pada tahun 2021, Kurniawan dan kawan-kawan melakukan penelitian mengenai klasterisasi tingkat Pendidikan di kecamatan yang ada di DKI Jakarta menggunakan algoritma K-Means[6].

Selain algoritma K-means dikenal juga algoritma K-Medoids yang juga termasuk kedalam algoritma pada metode berbasis partisi. K-Medoids adalah algoritma yang mengatasi kelemahan dari algoritma K-Means yang sensitive terhadap outliers[7]. Nurlaela dan kawan-kawan pada tahun yang sama juga melakukan penelitian klasterisasi penyakit juga khususnya *clustering* penyakit maag menggunakan algoritma K-Medoids[8]. Masih pada tahun yag sama, Sindi dan kawan-kawan juga melakukan penelitian tentang algoritma K-Medoids tetapi untuk klasterisasi penyebaran COVID-19 di Indonesia[9]. Pada tahun 2021 dilakukan penelitian oleh Sulistyawati dan Sadikin mengenai klasterisasi pelanggan menggunakan Algoritma K-Medoids[10].

Berdasarkan beberapa penelitian yang telah dilakukan dalam bidang klasterisasi yang telah dijabarkan diatas terlihat bahwa Algoritma berbasis partisi seperti K-Means dan K-Medoids banyak diimplementasikan pada penambangan pengetahuan di berbagai dataset berbeda. Hasil klasterisasi pada algoritma kedua algoritma ini bergantung pada penentuan jumlah k *cluster* awal yang biasanya ditentukan secara random.

Oleh sebab itu, pada penelitian ini mengimplementasikan algoritma *partitioning around medoid* untuk mengatasi masalah *outliers* pada dataset dan menggunakan metode *elbow* untuk menentukan jumlah *cluster* optimal. Untuk menguji kelayakan dan kualitas klasterisasi algoritma *partitioning around medoids* menggunakan metode *silhouette* dan metode *davies-bouldin indexs.* Tujuannya adalah untuk mengetahui seberapa jauh *cluster-cluster* terpisah dan seberapa padat *cluster-cluster* tersebut.

CLUSTERING

Teknik *clustering* ditemukan oleh Lloyd pada tahun 1957, metode *clustering* yang ditemukan oleh Llioyd adalah *K-Means clustering.* Pada algoritma *clustering* komputer akan mengelompokkan sendiri dataset yang menjadi inputan tanpa mengetahui target *class*-nya terlebih dahulu. Dataset akan dikelompokkan ke dalam *cluster* berdasarkan kemiripan dengan *instance* yang lain. Beberapa metode *clustering* yang sudah dikembangkan antara lain[11]:

1. *Partitional clustering*
   Exclusive clustering atau dikenal juga dengan *partitional clustering* merupakn jenis *clustering* yang dimana elemen-elemennya hanya dimiliki oleh sebuah *cluster* yang tidak boleh dimiliki oleh *cluster* lain.
2. *Overlapping clustering*
   Overlapping clustering disebut juga *soft clustering* yang merupakan jenis *clustering* yang elemen-elemennya boleh dimiliki oleh beberapa *cluster*.
3. *Hierarchical clustering*
   Hierarchical clustering disebut juga *multilevel hierarchy* yang mengelompokkan *cluster* yang lebih besar menjadi dua atau lebih *cluster* yang lebih kecil sehingga membentuk diagram pohon(*tree diagram).*
4. *Density based clustering*
   Density based clustering merupakan jenis *cluster* yang berkaitan dengan kerapatan objek,dimana *cluster* yang lebih padat dipisahkan oleh *cluster* yang lebih renggang.
5. *Model based clustering*

DOI: https://doi.org/10/25047/jtit.v8i2.258

Model based clustering merupakan jenis *clustering* yangelemennya dibentuk melalui asumsi atau model matematika atau model statistika standar.

K-MEDOIDS

Algoritma *K-medoids* termasuk kedalam metode *clustering* yang berbasis partisi (*Partitional clustering*) sama halnya dengan metode *K-Means* yang juga berbasis partisi. Algoritma *K-Medoids* mempartisi data atau membagi data kedalam kelompok-kelompok berbasis objek representasif (perwakilan). Algoritma *K-Medoids* merupakan algoritma yang digunakan untuk mengatasi kelemahan dari algoritma *K-Means*. Algoritma K-medoids memperbaharui centroids dengan objek aktual sebagai representasi dari suatu *cluster* bukan penggunaan rata-rata seperti dalam algoritma *K-Means*. Jadi algoritma *K-Medoids* meminimalkan jumlah perbedaan antara tiap objek p dan objek representasi terdekat ,menggunakan jumlah kesalahan absolut[7]:

$$E = \sum_{i=1}^{k} \sum_{p=c_i} dist(p, o_i) \quad (1)$$

Dimana E adalah jumlah kesalahan absolut untuk semua objek p dalam himpunan data . dan o_i adalah objek representative dari klaster C_i[7].

K-medoids diimplementasikan menggunakan Algoritma *Partitioning Around Medoid* (PAM). Langkah-langkah dalam algoritma *Partitioning Around Medoid* (PAM) adalah [7]:

1. Pilih sejumlah k objek dari himpunan dataset sebagai medoids awal
2. Periksa untuk semua kemungkinan objek non-representatif, apakah penggantian sebuah objek representative dengan objek non-representatif akan meningkatkan kualitas klasterisasi.
3. Ulangi Langkah ke dua hingga konvergen (Tidak dapat lagi meningkatkan kualitas klasterisasi) atau dengan kata lain hingga tidak ada lagi perubahan objek presentative.

METODE ELBOW

Metode *elbow* merupakan salah satu metode yang biasa digunakan untuk menentukan jumlah *cluster* terbaik dalam *clustering*. Analisis metode *elbow* dalam penentuan jumlah *cluster* terbaik dengan melihat bentuk siku (*elbow*) di dalam kurva yang dihasilkan yaitu perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik tertentu dalam kurva. Tahap-tahap dalam metode *elbow* antara lain[11]:

1. Menghitung jumlah WCSS (*within clusters sum of squares*) untuk beberapa nilai k yang ditentukan. WCSS ini berkaitan dengan jarak antara sampel yang menjadi elemen *cluster* dengan *centroid* nya
2. Gambarkan ke dalam grafik k dengan WCSS. Biasanya setiap pertambahan nilai k maka nilai WCSS akan menurun.

Formula untuk menentukan nilai WSS[11]:

$$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad (2)$$

METODE *SILHOUETTE*

Metode *silhouette* termasuk kedalam metode intrinsik yang akan mengevaluasi kualitas sebuah *clustering* dengan menguji seberapa baik *cluster* dipisahkan dan seberapa padat *cluster* tersebut. Biasanya metode intrinsik ini digunakan Ketika tidak terdapat klasterisasi yang ideal sebagai acuan. Koefisian *silhouette* menjadi ukuran dalam metode intrinsik, Adapun formulanya adalah[12]:

$$a(o) = \frac{\sum_{o' \in C_{i, o \neq o'}} dist(o, o')}{|C_i| - 1} \quad (3)$$

Dan

$$b(o) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{o' \in C_j} dist(o, o')}{|C_j|} \right\} \quad (4)$$

Serta silhouette coefficient dari o adalah[12]:

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

Nilai a(o) adalah kepadatan *cluster* yang mengandung objek o. Semakin kecil nilainya, semakin padat *cluster* tersebut. Nilai b(o) menangkap sejauh mana o dipisahkan dari yang *cluster* lain. Semakin besar b(o), semakin terpisah o dari *cluster* lain. Nilai koefisien *silhouette* adalah antara -1 dan 1.Oleh karena itu, Ketika nilai koefisien *silhouette* o mendekati 1, *cluster* yang berisi objek o sangat padat dan o jauh dari *cluster* lain[12].

Interpretasi subjektif dari koefisien *silhouette* yang didefinisikan sebagai lebar siluet rata-rata maksimal untuk seluruh kumpulan data ditunjukkan pada table 3[13].

TABEL I. PENGUKURAN KOEFISIEN SILHOUETTE

koefisien *silhouett*	**interpretasi yang diusulkan**
0,71-1	Struktur kuat
0,51-0,70	Struktur baik
0,26-0,50	Struktur lemah
$\leq 0,25$	tidak ada struktur substansial yang ditemukan

DAVIES-BOULDIN INDEX

Indeks davies-bouldin ini didasarkan pada gagasan bahwa untuk kebaikan partisi pemisahan antar *cluster* serta intra *cluster* homogenitas dan kepadatan. Kemudian, untuk menentukan nilai *davies-bouldinm index*, perlu menentukan ukuran dispersi dan ukuran kesamaan *cluster*. Dalam dispersi S_i cluster C_i dan pemisahan D_{ij} antara *i*th dan *j*th *cluster* didefinisikan sebagai[14]:

$$S_i = \left(\frac{1}{|C_i|} \sum_{x \in C_I} D^p(x, C_i) \right)^{\frac{1}{p}}, p > 0 \quad (6)$$

Dimana $|C_i|$ adalah jumlah titik data dalam *cluster* C_i. Dan c_i adalah pusat *cluster* C_i[13]:

$$D_{ij} = \left(\sum_{l=1}^{d} |v_{il} - v_{jl}|^t \right)^{\frac{1}{t}}, t > 1 \quad (7)$$

DOI: https://doi.org/10/25047/jtit.v8i2.258 ©2021 JTIT

Dimana v_{il} dan v_{jl} adalah *centroid* dari *cluster* D_i dan D_j. Dan nilai *davies bouldin index* di rumuskan sebagai berikut[13]:

$$V_{DB} = \frac{1}{4}\sum_{i=1}^{k} R_i \quad (8)$$

Dimana k adalah jumlah *cluster* dan R_i adalah[13]:

$$R_i = \underbrace{max}_{j \neq i} R_{ij} \quad (9)$$

R_{ij} adalah ukuran kesamaan antar *cluster* C_i dan C_j.formula R_{ij} adalah sebagai berikut[13]:

$$R_{ij} = \frac{S_i - S_j}{D_{ij}} \quad (10)$$

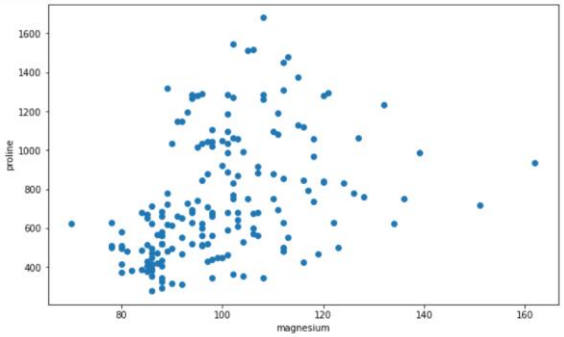
Karena tujuannya adalah untuk mencapai minimum within-cluster dan maksimum *within-cluster dispersion* dan pemisahan antar *cluster* maksimum, jumlah *cluster[14]*.

Semakin rendah nilai *davies-bouldin indexs* semakin baik kualitas *cluster* yang diperoleh[15].

METODE PENELITIAN

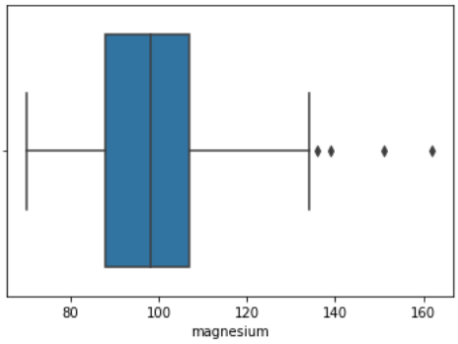
A. Analisis Data

Dataset yang digunakan pada penelitian ini di ambil dari dataset *UCI repository* yaitu dataset wine. Menurut informasi dari 'UCI Machine Learning Repository' bahwa diketahui dataset *wine* memiliki 13 atribut. Visualisasi data yang ditampilkan untuk atribut magnesium dan proline seperti pada Gambar 1.



Gambar. 1 Plot mengenai magnesium yang dimiliki wine terhadap kandungan kandungan proline

Berdasarkan perhitungan nilai *Interquartile Range(IQR)* pada nilai atribut magnesium [16], dimana nilai maksimum data lebih besar dari IQR maksimum, sehingga terdapat *high outlier*. IQR minimum = -0.11413043478260865 dan IQR maksimum = 0.7119565217391306 sedangkan nilai minimum data = 0.0 dan nilai maksimum data = 1.0. Gambar 2 menunjukkan *outliers* pada atribut magnesium dalam dataset *wine.*

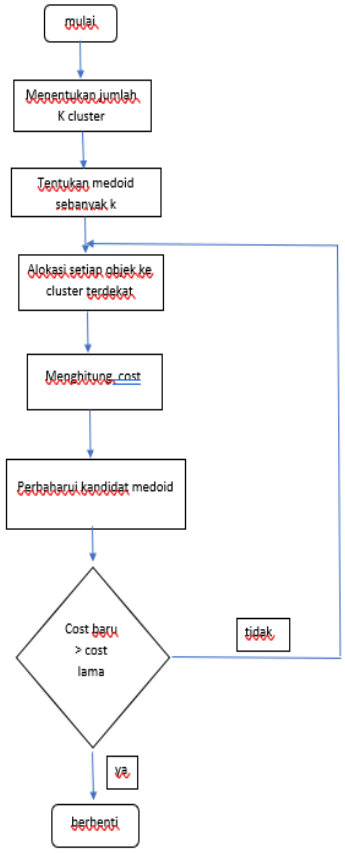


Gambar. 2 grafik outliers pada atribut magnesium

B. Penemuan Pola

Pada algoritma K-*Medoids* Langkah-langkah untuk penemuan pola nya tidak jauh berbeda dengan Langkah-langkah pada algoritma K-Means seperti penenetuan jumlah k *cluster*. Algoritma *K-Medoids* bergantung sekali dengan penetapan nilai k yang selama ini dipilih secara random(acak) dengan cara *try* and *error*.

Langkah pertama dalam algoritma K-Medoids adalah menentukan medoid untuk sejumlah k *cluster* secara acak. Kemudian Langkah kedua adalah kelompokkan semua objek yang bukan medoid dalam dataset ke *cluster* yang memiliki objek medoid terdekat diantara k objek medoid yang ada menggunakan fungsi biaya. Lakukan pengecekkan untuk semua kemungkinan objek yang bukan medoid.ulangi Langkah tersebut sampai tidak ada lagi perubahan dalam medoid (pusat *cluster*)[7]. Flowchart Algoritma K-Medoids ditunjukkan pada Gambar 3.

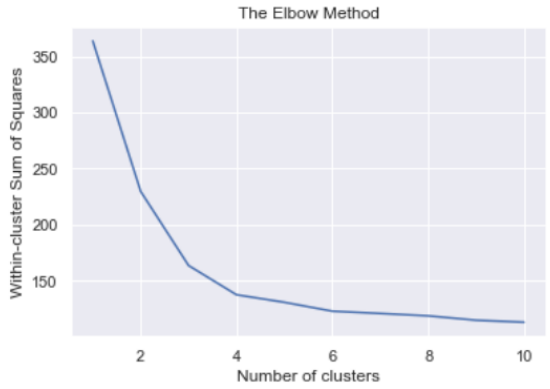


Gambar. *3 Flowchart* Algoritma *K-Medoids*

HASIL DAN PEMBAHASAN

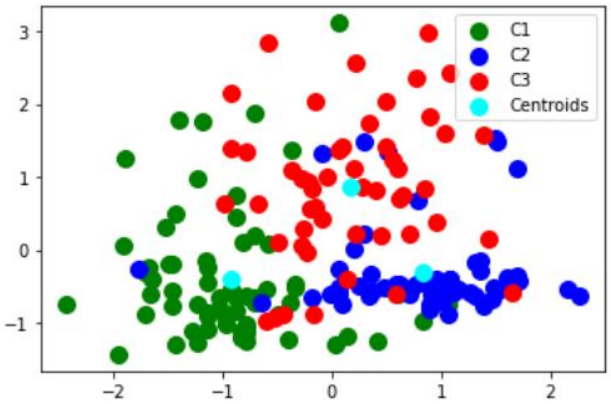
A. Analisis metode *Elbow*

Berdasarkan hasil perhitungan *within-cluster sum of squares* atau WCSS dari k=1 hinggan k= 10 terlihat ada tiga bentuk siku yang terbentuk pada grafik di k=2 , di k=3 dan k=4. Penjelasan tersebut ditunjukkan pada Gambar 4.



Gambar. 4 Penentuan jumlah k dengan metode elbow

Terbentuknya siku tersebut dijadikan indikator sabagai jumlah k terbaik. Akan tetapi ada 3 bentuk siku yang terlihat dalam gambar grafik pada Gambar 2. Tiga siku tersebut ada di k=2 , di k=3 dan k=4. Diantara ke tiga siku yang terlihat dalam grafik pada Gambar 4 untuk nilai k *cluster* optimalnya adalah k=3 .Alasan mengapa ditentukan jumlah k optimal adalah 3 karena terlihat penurunan nilai yang cukup besar dari titik k = 3 ke titik k = 4. Sedangkan pada titik k =4 ke titik k = 5 penurunan nilainya tidak sebesar k = 3 ke k = 4.Visualisasi data untuk model klasterisasinya di tunjukkan pada Gambar 5.



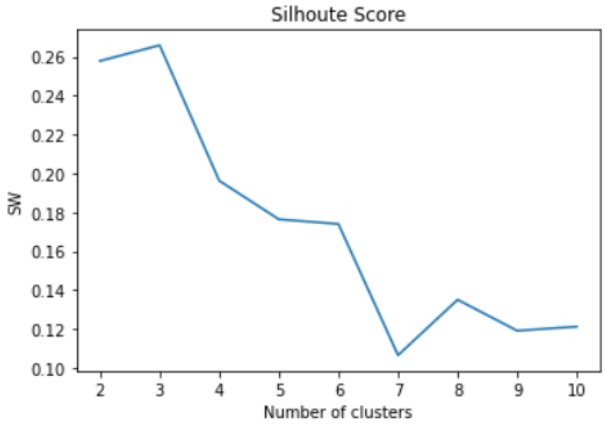
Gambar. 5 Hasil klasterisasi metode *elbow*

B. Evaluasi menggunakan metode Silhouette

Uji kelayakan dan kualitas klasterisasi yang dihasilkan oleh algoritma K-Medoids dengan penentapan k= 3 berdasarkan nilai koefisien *silhouette* adalah dimana k= 3 nilai koefisien *silhouette* nya adalah 0,27. Gambar 6 menunjukkan grafik *silhouette*. Terlihat pada hasil perbandingan nilai koefisien *silhouette* pada tabel 2 dari k=2 hingga k=10 bahwa pada nilai k= 3 adalah nilai koefisien *silhouette* paling besar mendekati 1 daripada nilai k yang lain. Tetapi Berdasarkan koefisien *silhouette* menurut Kauffman dan Roesseeuw rentang nilai yang dihasilkan tersebut termasuk kriteria struktuk lemah.

TABEL II. NILAI KOEFISIEN SILHOUETTE

Jumlah cluster	Silhouette Coefficient (pembulatan)
2	0,26
3	0,27
4	0,20
5	0,18
6	0,17
7	0,11
8	0,14
9	0,12
10	0,12



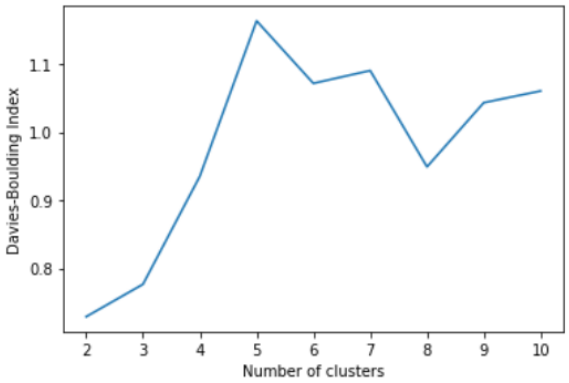
Gambar. 6 Penentuan jumlah k dengan metode silhoutte

C. Evaluasi menggunakan metode *davies-bouldin indexs*

Menurut evaluasi metode *davies bouldin indexs* bahwa nilai *davies bouldin indexs* untuk jumlah *cluster* k = 3 adalah 0,776.

TABEL III. NILAI INDEX DAVIES-BOULDIN

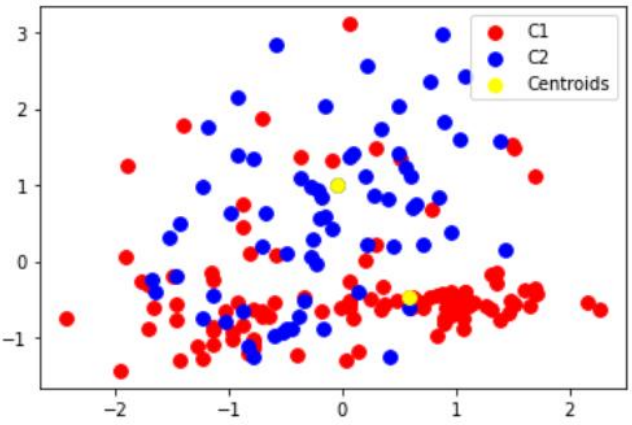
Jumlah cluster	Index davies-bouldin (pembulatan)
2	0,729
3	0,776
4	0,935
5	1,164
6	1,072
7	1,091
8	0,949
9	1,043
10	1,060



Gambar. 7 Grafik nilai *davies bouldin indexs* untuk tiap kelasS

DOI: https://doi.org/10/25047/jtit.v8i2.258 ©2021 JTIT

Tetapi k =3 bukanlah klasterisasi yang optimal menurut perhitungan nilai *davies bouldin indexs* pada table 3 maka diperoleh model dengan pemisahan paling baik antara cluster adalah di k = 2. Yaitu sebesar 0,729. Gambar 7 memperlihatkan grafik hasil perhitungan *davies-bouldin indexs* untuk k = 2 hingga k = 10.



Gambar. 7 Hasil klasterisasi metode *davies bouldin indexs*

KESIMPULAN

Algoritma K-Medoids dapat melakukan klasterisasi terhadap dataset yang memiliki *outliers*. Berdasarkan hasil kasterisasi diperoleh bahwa dataset *wine* paling optimal dibagi menjadi 3 *cluster*. Metode *elbow* dapat membantu untuk menetapkan jumlah k *cluster* terbaik. Sehingga dalam penetapan jumah k *cluster* tidak perlu lagi secara manual (try and error).

Berdasarkan evaluasi pengukuran kualitas klasterisasi menggunakan metode *silhouette*, menunjukkan bahwa *cluster* yang dihasilkan algoritma ini adalah termasuk struktur lemah. Sehingga isi dari setiap *cluster* tidak padat dan objek antar *cluster* memiliki jarak yang dekat(tumpeng tindih).

Sedangkan hasil analisis menurut perhitungan metode *davies-bouldin indexs* menunjukkan bahwa dataset *wine* paling optimal di bagi kedalam 2 *cluster* .

PENGHARGAAN

Terimakasih kepada pihak LPPM ITN Malang yang sudah bersedia mendanai penelitian ini.

REFERENSI

[1] Gustientiedina.,Hasnul,M,A.,Desnelita, Y.(2019). Penerapan Algoritma K-Means untuk clustering data Obat-Obatan pada RSID Pekanbaru.*Jurnal Nasional teknologi dan Sistem Informasi*,5(1),017-024.ISSN(online):2476-8812. https : //doi.org/10.25077/ TEKNOSI.v5i1.2019.17-24

[2] Annur.Haditsah.(2019).Penerapan data Mining Menentukan Strategi penjualan Variasi mobil menggunakan Metode K-Means Clustering(Studi kasus Toko Luxor Variaso Gorontalo).*Jurnal Informatika UPGRIS*,5(1),2447-6645.E-ISSN: 2460-4801.

[3] Mustofa.(2019).Penerapan Algoritma K-Means Clustering pada Karakter Permainan Multiplayer Online battle Arena.*Jurnal Informatika*,6(2),246-254.E-ISS:2528-2247

[4] Siringoringo.Rimbun.,Jamaluddin.,Perangin Angin.R.(2020).Pemodelan topik berita menggunakan Latent Dirichlet Allocation dan K-Means Clustering.*Jurnal Informatika Kaputama(JIK)*.4(2).E-ISSN:2685-5240.

[5] Hamzaoui.Y.,Amnai.M.,Choukri.A.,Fakhri.Y.(2020).Enhancenig OLSR Routing Protocol Using K-Means Clustering in MANETs.*International Journal of Electrical and Computer Engineering(IJECE)*,10(4), 3715-3724.ISSN:2088-8708. https://doi.org/10.25077/ TEKNOSI.v5i1.2019.17-24.

[6] Kurniawan.Ragil.,Mukarrobin.M.,Mahradianur.(2021).Klasterisasi Tinglat Pendidikan di DKI Jakarta pada Tinglat Kecamatan Menggunakan Algoritma K-Means.*Technologia*,12(4),234-239.

[7] Suyanto.(2019).data Mining untuk Klasifikasi dan Klasterisasi data edisi revisi.penerbit: Informatika Bandung.ISBN:978-602-6232-97-7.

[8] Nurlaela.Siti.,Primajaya.Aji.,Padilah.Tesa Nur.(2020).Algoritma K-Medoids untuk Clustering Penyakit Maag di kabupaten karawang.*Jurnal Informatika, Manajemen dan Komputer* ,12(2).eISSN: 2580-3042

[9] Sindi.Sukma.,Ningse.,weni ratnasari Orktapia,Sihombing.,Irma Agustika.Zer.,Fikrul Ilmi R.H, Hartama.,Dedy.(2020)Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran COVID-19 di Indonesia.*Jurnal teknologi Informasi*,4(1).E-ISSN: 2615-2738.

[10] Sulistyawati.Anggi Ayu Dwi.,sadikin.Mujiono.(2021).Penerapan Algoritma K-Medoids untuk menentukan Segmentasi Pelanggan. *Jurnal Sistem Informasi(SISTEMASI)*,10(3),516-526.e-ISSN:2540-9719.

[11] Purnama.Bedy.(2019).Pengantar Machine Laerning Konsep dan Praktikum dengan Contoh Latihan Berbasis R dan Python.penerbit:Informatika Bandung.ISBN:978-623-7131-19-9.

[12] Han.Jiawei.(2012).Data Mining concept and Techniques.Third edition. Morgan Kaufmann Publisher is an Imprint of Elsevier.USA.

[13] Kaufman.Leonard.,Rousseeuw.Peter J.(2005).Finding Groups in Data(an Introduction to Cluster Analysis).Publish by Jonh Wiley dan Sons,Inc,Hoboken,New Jersey.Published simultraneosly in Canada.ISBN: 0-471-73578-7

[14] Thomas.,Juan Carlos Rojas.,Cofre.Marco Mora.,Santos.Matilde.(2014). New Version of Davies-Bouldin index for clustering validation based on hyper rectangles.*Conference Paper*.*January 2014. DOI: 10.1049/14.2014.0001*.

[15] Dewi.Dewa Ayu Indah Cahya., Pramita.Dewa Ayu Kadek.(2019).Analisis perbandingan Metode elbow dan Silhouette pada Algoritma Clustering K=Medoids dalam Pengelom[okan Produksi Kerajinan Bali.*Jurnal Matrix*,9(3).