e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 5 No. 1, Juni, 2021, **Hal. 99-107**
DOI: 10.29408/edumatic.v5i1.3359

Komparasi *Distance Measure* Pada K-Medoids *Clustering* untuk Pengelompokkan Penyakit Ispa

Mia Nuranti Putri Pamulang[*1], Mia Nuur Aini[2], Ultach Enri[3]
[1,2,3] Program Studi Teknik Informatika, Universitas Singaperbangsa Karawang
email: mia.nuranti17129@student.unsika.ac.id*[1], mia.nuuraini17130@student.unsika.ac.id[2], ultach@staff.unsika.ac.id[3]

(Received: 25 April 2021/ Accepted: 31 Mei 2021 / Published Online: 20 Juni 2021)

Abstrak

K-Medoids adalah algoritma *unsupervised* yang menggunakan *distance measure* untuk menglompokkan data. *Distance measure* merupakan metode pengukuran jarak yang dapat membantu sebuah algoritma mengelompokkan objek berdasarkan kemiripan variabel-variabelnya. Beberapa penelitian telah menunjukkan bahwa penggunaan *distance measure* yang tepat dapat meningkatkan performa algoritma dalam melakukan klastering. *Euclidean* dan *Chebyshev* adalah dua dari beberap *distance measure* yang dapat digunakan. Pada tahun 2016, Dinas Kesehatan Karawang menyatakan sebanyak 175.891 warga Karawang menderita penyakit ISPA Angka ini terus bertambah pada tahun berikutnya hingga pada tahun 2019. Total warga Karawang yang menderita penyakit ISPA mencapai 181.945. Untuk membantu pemerintah dalam menanggulangi masalah ini, maka akan dilakukan klastering untuk mengelompokkan daerah penyebaran penyakit ISPA di Kabupaten Karawang. Daerah penyebaran penyakit ISPA akan dibagi menjadi tiga kelompok yaitu rendah, sedang dan tinggi. Komparasi *distance measure* dilakukan untuk menemukan model terbaik berdasarkan evaluasi *Davies Bouldin Index* (DBI). Penggunaan *Euclidean distance* menghasilkan nilai DBI sebesar 0,088 sedangkan penggunaan *Chebyshev distance* menghasilkan nilai DBI sebesar 0,116. Performasi algoritma K-Medoids dengan *Euclidean distance* dianggap lebih baik dari *Chebyshev distance* karena menghasilkan nilai DBI yang mendekati 0.

Kata kunci: K-Medoids, *Euclidean Distance, Chebyshev Distance, Davies Bouldin Index*

Abstract

K-Medoids is an unsupervised algorithm that uses a distance measure to classify data. The distance measure is a method that can help an algorithm classify data based on the similarity of the variables. Several studies have shown that using the right distance measure can improve the performance of the algorithm in clustering. Euclidean and Chebyshev is two of some distance measures that can be used. In 2016, Karawang Health Office stated that 175.891 Karawang citizens were suffering from ISPA. This figure continued to increase in the following year until 2019. The total of Karawang citizens who suffering from ISPA reached 181.945 people. To assist the government in overcoming this problem, a clustering process will be carried out to group the areas where the ISPA is spreading in Karawang District. The area will be divided into three clusters, namely low, medium and high. Comparison of distance measures is carried out to find the best model based on the evaluation of the Davies Bouldin Index (DBI). The use of Euclidean-distance produces a DBI score of 0,088 meanwhile the use of Chebyshev distance resulted in a DBI score of 0,116. The performance of the K-Medoids algorithm with Euclidean-distance is considered to be better than Chebyshev distance because it produces a DBI score that is near to 0.

Keywords: *K-Medoids, Euclidean Distance, Chebyshev Distance, Davies Bouldin Index*

PENDAHULUAN

Data Mining merupakan suatu proses penggalian informasi melalui sekumpulan data dengan tujuan memperoleh pengetahuan untuk diterapkan pada bidang-bidang tertentu. Dalam praktiknya, kita dapat memanfaatkan data mining untuk melakukan klasifikasi (Sari, Firdausi, & Azhar, 2020), prediksi, estimasi, klastering maupun asosiasi. Subash pada penelitiannya di tahun 2016 menyebutkan beberapa hal yang dapat dilakukan data mining pada dunia medis diantaranya ; mendeteksi penipuan asuransi kesehatan, memberikan solusi medis yang lebih baik untuk pasien yang kekurangan biaya, membantu mendeteksi penyakit tertentu, dan mengidentifikasi metode perawatan medis yang efisien (Pandey & Jain, 2017). Selain itu, data mining diartikan sebagai proses pencarian pola berupa pengetahuan yang tersembunyi (Takdirillah, 2020), dengan menggunakan teknik statistic, kecerdasasan buatan, dan machine learning untuk menemukan pengetahuan langsung pada database (*knowledge discovery in database*)(de la Vega, García-Saiz, Zorrilla, & Sánchez, 2020; Ishak, Siregar, Ginting, & Afif, 2020). Data mining juga dapat membantu mengetahui wilayah penyebaran penyakit pada suatu daerah dengan metode klastering.

Infeksi Saluran Pernapasan Akut atau yang biasa disebut ISPA adalah salah satu penyakit yang disebabkan oleh bakteri atau virus. Penyakit ini menyerang organ tubuh di bagian saluran pernafasan bagian atas maupun bawah dan akan menimbulkan gejala batuk, pilek disertai dengan demam. Adapun beberapa kelompok orang yang rentan tertular penyakit ini berdasarkan yang disebutkan Alodokter, yaitu anak-anak dan lansia, orang dengan sistem kekebalan tubuh lemah, penderita gangguan jantung dan paru-paru, serta perokok aktif.

Pada tahun 2016, Dinas Kesehatan Karawang menyatakan sebanyak 175.891 warga Karawang menderita penyakit ISPA Angka ini terus bertambah pada tahun berikutnya hingga pada tahun 2019, total warga Karawang yang menderita penyakit ISPA mencapai 181.945. Untuk membantu pemerintah dalam menanggulangi masalah ini, maka akan dilakukan klastering untuk mengelompokkan daerah penyebaran penyakit ISPA di Kabupaten Karawang. Klastering adalah sebuah metode *unsupervised* yang dapat membagi objek menjadi beberapa kelompok berdasarkan kemiripannya tanpa harus dilakukan pelatihan terlebih dahulu. Pada klastering, hasil yang baik adalah jika setiap kelompok memiliki kemiripan yang tinggi antar objek (Ningrat, Maruddani, & Wuryandari, 2016). Metode ini digunakan oleh Ade Bastian dan kawan-kawan dalam penelitiannya di tahun 2018. Tri Juninda dan kawan-kawan menggunakan metode ini untuk mengelompokkan penyakit berdasarkan daerah tertentu menggunakan algoritma K-Medoids.

Kedua penelitian yang telah disebutkan menggunakan algoritma yang berbeda saat melakukan klastering. Ade Bastian dan kawan-kawan menggunakan algoritma K-Means dalam penelitiannya karena algoritma ini dinilai memiliki ketelitian yang tinggi terhadap ukuran objek. Saat mengolah data dalam jumlah besar, K-Means dianggap relatif lebih terukur dan efisien (Bastian, Sujadi, & Febrianto, 2018). Sedangkan pada penelitian yang dilakukan Tri Juninda dan kawan-kawan, K-Medoids dianggap memiliki karakteristik penting dibanding K-Means karena dapat menghitung *medoids* menggunakan frekuensi kemunculannya (Juninda, Mustasim, & Andri, 2019). K-Medoids cukup efisien untuk mengolah data dalam jumlah kecil serta mampu mengatasi kelemahan K-Means yang sensitif terhadap *outlier* (Irawan, Siregar, Damanik, & Saragih, 2020).

Saat melakukan klastering, *distance measure* pada algoritma yang digunakan akan sangat berpengaruh pada hasil yang didapatkan (Miftahuddin, Umaroh, & Karim, 2020). *Distance measure* merupakan sebuah metode pengukuran jarak antar objek (Liu, Zhang, Zhang, & Cui, 2017; Tao et al., 2019). Semakin besar nilai jarak yang diperoleh, maka semakin jauh letak objek dengan pusat klaster yang terbentuk (Nahdliyah, Widiharih, & Prahutama, 2019). Beberapa *distance measure* yang dapat digunakan antara lain ; *Euclidean Distance*, *Chebyshev Distance*, *Manhattan Distance*, dan *Minkowski Distance* (Gueorguieva,

Valova, & Georgiev, 2017; Gultom, Sriadhi, Martiano, & Simarmata, 2018; He, Agard, & Trépanier, 2020; Saputra, Saputra, & Oswari, 2020).

Berbagai penelitian mengenai penerapan *distance measure* telah dilakukan pada tahun 2019, Nishom melakukan perbandingan akurasi pada algoritma K-Means dengan beberapa *distance measure* yaitu *euclidean, minkowski,* dan *manhattan* kemudian menyimpulkan bahwa penggunaan *euclidean distance* menghasilkan tingkat akurasi tertinggi dibanding dua *distance measure* lainnya (Nishom, 2019). Penelitian lain dilakukan oleh (Mustofa & Suasana, 2018) mengenai penerapan K-Medoids untuk penentuan status EDGI (*E-Goverment Development Index*). Mereka menggunakan *chebyshev* sebagai *distance measure* dan *Davies Bouldin Index* sebagai validasi hasil akhir penelitiannya. Penelitian ini menyimpulkan bahwa *Chebyshev distance* pada K-Medoids berhasil mengoptimalkan penentuan pengelompokkan EDGI jika dibandingkan dengan *manhattan* dan *euclidean distance*

Berdasarkan hal tersebut di atas, pada penelitian ini menggunakan K-Medoids untuk melakukan klastering serta membandingkan dua *distance measure* yaitu *euclidean* dan *chebyshev* dengan validasi *Davies Bouldin Index* (DBI). Adapun objek yang diteliti merupakan dataset penyakit ISPA di Kabupaten Karawang tahun 2019. Metodologi yang digunakan adalah *Knowledge Discovery in Database* (KDD) dengan bantuan *tool* RapidMiner.

METODE

Penelitian ini dilakukan dengan metodologi *Knowledge Discovery in Database* (KDD). KDD sendiri merupakan proses menemukan pengetahuan dalam kumpulan data (Ghazal & Hammad, 2020; Schmidt & Sun, 2018) yang diberikan terlepas dari karakteristik dan ukuran atribut didalamnya (Kumar, Jain, & Chauhan, 2019). Menurut mereka, beberapa langkat=h untuk memahami dan mengekstraksi pola dari kumpulan data yaitu *data selection, data preprocessing, data transformation, data mining,* dan *interpretation/ evaluation*. Adapun instrumen yang digunakan sebagai pendukung dalam penelitian ini yaitu berupa data dokumentasi dari Dinas Kesehatan Kesehatan Karawang dan data tersebut diolah dengan bantuan *tool* RapidMiner. RapidMiner adalah aplikasi yang digunakan untuk memperoleh informasi dan pengetahuan pada data dengan menganalisis kuantitas data secara kualitatif (Uska, Wirasasmita, Usuluddin, & Arianti, 2020).

Data yang digunakan pada penelitian ini adalah dataset penyakit ISPA yang didapat dari Dinas Kesehatan Karawang tahun 2019. Setelah itu dataset yang didapat akan melewati tahap seleksi atribut dimana atribut yang dipilih adalah atribut yang relevan dengan penelitian ini. Setelah itu dataset tersebut akan ditransformasi dengan metode *min-max normalization* untuk menyetarakan *range* setiap variabelnya menjadi bernilai 0 sampai 1 (Santoso, Cholissodin, & Setiawan, 2017).

Langkah selanjutnya adalah melakukan klastering dengan algoritma K-Medoids. Algoritma ini merupakan algoritma pengelompokkan yang menggunakan objek data sebagai perwakilan (*medoids*) sebagai pusat klaster dengan meminimalkan jumlah kesamaan antar setiap objek dan titik referensi yang sesuai (Gunawan, Anggraeni, Rini, & Mustofa, 2020). Saat melakukan klastering dengan algoritma K-Medoids, akan dilakukan 2 kali percobaan dengan *distance measure* yang berbeda yaitu *Euclidean* dan *Chebyshev*.

Setelah mendapatkan hasil klastering dengan masing-masing *distance measure*, langkah terakhir adalah melakukan evaluasi hasil klaster dengan menggunakan *Davies Bouldin Index* (DBI) dimana DBI akan mengukur kedekatan antar data dalam satu kelompok data dengan menghitung standar deviasinya (Nawrin, Rahatur, & Akhter, 2017). Adapun *range* nilai *Davies Bouldin Index* adalah 0 sampai 1. Semakin kecil nilai yang didapat, maka kemiripan data dalam satu kelompok akan semakin tinggi. Nawrin dan kawan-kawan

menjelaskan bahwa nilai DBI terkecil dianggap sebagai algoritma terbaik berdasarkan kriteria ini.

HASIL DAN PEMBAHASAN

Hasil

Penelitian ini dilakukan untuk melakukan perbandingan dua *distance measre* yang digunakan pada algoritma K-Medoids. Terdapat 25 atribut pada dataset yang digunakan dan diseleksi 9 atribut dengan relevansi tinggi untuk mendukung penelitian. Seleksi atribut dilakukan secara manual dengan bantuan *tool* Ms.Excel. Adapun 9 dari 25 atribut yang telah diseleksi akan ditampilan pada Tabel 1.

Tabel 1. Hasil Seleksi Atribut

No	Atribut
1	Nama Puskesmas
2	Jumlah Penduduk
3	Pneumonia < 1 Tahun
4	Pneumonia 1 - < 5 Tahun
5	Pneumonia ≥ 5 Tahun
6	Total Pneumonia
7	Total Bukan Pneumonia < 1 Tahun dan 1 - < 5 Tahun
8	Bukan Pneumonia ≥ 5 Tahun
9	Total Bukan Penumonia

Pada Tabel 1, Nama Puskesmas adalah atribut yang akan dijadikan label klastering. Selebihnya adalah atribut jumlah penderita Pneumonia dan Bukan Pneumonia berdasarkan usianya yang akan dijadikan variabel untuk perhitungan klastering menggunakan *Euclidean* dan *Chebyshev Distance.* Dari seluruh fitur yang diseleksi, 8 diantaranya akan dilakukan transformasi data dengan metode *min-max normalization*.

Tabel 2. Dataset Sebelum Normalisasi

Puskesmas	JP	P < 1	P 1 - < 5	P > 5	TP	TBP < 1 & 1 - < 5	BP > 5	TBP
Adiarsa	65962	20	83	3	106	651	1110	1761
Anggadita	24394	7	83	111	201	431	933	1364
Balongsari	19543	2	76	0	78	193	1042	1235
Batujaya	81806	4	29	0	33	1352	1321	2673
Bayur Lor	24847	0	0	0	0	1445	585	2030
Ciampel	38390	11	27	79	117	512	1064	1576
Cibuaya	53035	44	37	128	209	522	506	1028
Cicinde	30589	0	54	0	54	1457	2950	4407
Cikampek	111415	36	98	0	134	1591	2967	4558
…	…	…	…	…	…	…	…	…
Wanakerta	4902	0	0	0	0	388	6541	6929

Tabel 2 menampilkan *record* jumlah penderita Pneumonia maupun Bukan Pneumonia berdasarkan usia penderita sebelum dilakukan normalisasi. Pada tabel tersebut terlihat jelas rentang jarak nilai pada masing-masing atribut memiliki perbedaan yang sangat besar. Hal ini dapat mempengaruhi hasil dari proses data mining selanjutnya, maka dari itu dataset harus

dinormalisasi terlebih dahulu. Selanjutnya, tabel 3 menampilkan *record* jumlah penderita Pneumonia maupun Bukan Pneumonia berdasarkan usia penderita setelah melewati proses normalisasi data. Pada tabel tersebut dapat dilihat perbedaannya yaitu nilai pada masing-masing atribut memiliki rentang seimbang dengan hasil normalisasi terkecil adalah 0 dan terbesar adalah 1.

Tabel 3. Dataset Setelah Normalisasi

Puskesmas	JP	P < 1	P 1 - < 5	P > 5	TP	TBP < 1 & 1 - < 5	BP > 5	TBP
Adiarsa	0,505	0,018	0,076	0,021	0,048	0,128	0,170	0,234
Anggadita	0,503	0,006	0,076	0,787	0,092	0,085	0,143	0,181
Balongsari	0	0,002	0,069	0	0,036	0,038	0,159	0,164
Batujaya	0,678	0,004	0,026	0	0.015	0,265	0,202	0,355
Bayur Lor	0,058	0	0	0	0	0,284	0,089	0,270
Ciampel	0,205	0,010	0,025	0,560	0,053	0,100	0,163	0,209
Cibuaya	0,365	0,040	0,034	0,908	0,095	0,102	0,077	0,137
Cicinde	0,120	0	0,049	0	0,025	0,286	0,451	0,585
Cikampek	1	0,033	0,089	0	0,061	0,312	0,454	0,605
…	…	…	…	…	…	…	…	…
Wanakerta	0,321	0	0	0	0	0,076	1	0,920

Berikut adalah keterangan dari masing-masing atribut :

JP : Jumlah Penduduk
P < 1 : Pneumonia < 1 Tahun
P 1 - < 5 : Pneumonia 1 sampai < 5 Tahun
P > 5 : Pneumonia > 5 Tahun
TBP 1 & 1 - < 5 : Total Bukan Pneumonia 1 dan 1 sampai < 5 Tahun
BP > 5 : Bukan Penumonia > 5 Tahun
TBP : Total Bukan Pneumonia

Untuk membandingkan kedua *distance measure* yang telah disebutkan, maka dilakukan percobaan sebanyak 2 kali dimana percobaan pertama adalah melakukan klastering menggunakan K-Medoids dengan persamaan *Euclidean Distance*. Kemudian pada percobaan yang kedua akan dilakukan klastering menggunakan K-Medoids dengan persamaan *Chebyshev Distance.* Tabel 4 adalah anggota yang didapatkan pada masing-masing percobaan klastering.

Berdasarkan hasil klaster yang dapat dilihat pada tabel diatas, penggunaan masing-masing *distance measure* menghasilkan jumlah anggota klaster yang berbeda. Penggunaan *Euclidean Distance* menghasilkan 6 anggota pada klaster dengan daerah penyebaran penyakit ISPA yang rendah (*cluster_0*) yaitu daerah Cikampek Utara, Gempol, Kalangsari, Lemah Abang, Tirta Jaya dan Wanakerta. Kemudian menghasilkan 44 anggota pada klaster dengan daerah penyebaran penyakit ISPA yang sedang (*cluster_1)* yaitu daerah Adiarsa, Anggadita, Balongsari, Batujaya, Bayur Lor, Ciampel, Cibuaya, Cicinde, Cikampek, Cilamaya, Curug, Jatisari, Jayakerta, Jomin, Karawang, Karawang Kulon, Kerta Mukti, Kutamukti, Lemahdhuhur, Loji, Kutawaluya, Majalaya, Medang Asem, Kota Baru, Nagasari, Pacing, Pakis Jaya, Pangkalan, Pasirukem, Pedes, Plawad, Purwasari, Rawamerta, Rengas Dengklok, Sukatani, Tanjung Pura, Telaga Sari, Teluk Jambe, Sungai Buntu, Tempuran, Tirta Mulya,

Tunggak Jati, dan Wadas. Dan terakhir menyisakan 1 anggota pada klaster dengan daerah penyebaran penyakit ISPA yang tinggi (*cluster_2*) yaitu daerah Klari.z

Penggunaan *Chebyshev Distance* menghasilkan 4 anggota pada klaster dengan daerah penyebaran penyakit ISPA yang rendah (*cluster_0*) yaitu daerah Cikampek Utara, Gempol, Lemah Abang, dan Wanakerta. Kemudian menghasilkan 41 anggota pada klaster dengan daerah penyebaran penyakit ISPA yang sedang (*cluster_1*) yaitu daerah Adiarsa,Anggadita, Balongsari, Batujaya, Ciampel, Cicinde, Cikampek, Cilamaya, Curug, Jatisari, Bayur Lor, Jayakerta, Jomin, Kalangsari, Karawang, Karawang Kulon, Kerta Mukti, Medang Asem, Klari, Kutamukti, Kutawaluya, Lemahdhuhur, Loji, Majalaya, Nagasari, Pacing, Pakis , Jaya, Pangkalan, Pedes, Purwasari, Rengasdengklok, Pasirukem, Tanjung Pura, Telaga Sari, Tempuran, Tirta Mulya, Tunggak Jati, dan Wadas. Dan terakhir menyisakan 6 anggota pada klaster dengan daerah penyebaran penyakit ISPA yang tinggi (*cluster_2*) yaitu daerah Cibuaya, Kota Baru, Klawad, Rawa Merta, Sungai Buntu, dan Tirta Jaya.

Hasil klastering yang didapat selanjutnya dievaluasi dengan *Davies Bouldin Index* (DBI) untuk mengetahui performansi masing-masing *distance measure* pada algoritma K-Medoids. Tabel 5 adalah hasil nilai DBI pada masing-masing *distance measure* yang didapatkan dengan bantuan *tool* RapidMiner. Pada evaluasi *Davies Bouldin Index*, nilai yang mendekati 0 menunjukkan bahwa hasil klaster memiliki kemiripan yang tinggi dalam satu kelompoknya. Berdasarkan Tabel 5, *Euclidean Distance* mendapatkan nilai DBI sebesar 0,088 sedangkan *Chebyshev Distance* mendapatkan nilai DBI sebesar 0,116.

Tabel 4. Anggota Klaster

	Cluster_0	***Cluster_1***	***Cluster_2***
Euclidean Distance	Cikampek Utara, Gempol, Kalangsari, Lemah Abang, Tirta Jaya, Wanakerta	Adiarsa, Anggadita, Balongsari, Batujaya, Bayur Lor, Ciampel, Cibuaya, Cicinde, Cikampek, Cilamaya, Curug, Jatisari, Jayakerta, Jomin, Karawang, Karawang Kulon, Kerta Mukti, Kutamukti, Lemahdhuhur, Loji, Kutawaluya, Majalaya, Medang Asem, Kota Baru, Nagasari, Pacing, Pakis Jaya, Pangkalan, Pasirukem, Pedes, Plawad, Purwasari, Rawamerta, Rengas Dengklok, Sukatani, Tanjung Pura, Telaga Sari, Teluk Jambe, Sungai Buntu, Tempuran, Tirta Mulya, Tunggak Jati, Wadas	Klari
Chebyshev Distance	Cikampek Utara, Gempol, Lemah Abang, Wanakerta	Adiarsa,Anggadita, Balongsari, Batujaya, Ciampel, Cicinde, Cikampek, Cilamaya, Curug, Jatisari, Bayur Lor, Jayakerta, Jomin, Kalangsari, Karawang, Karawang Kulon, Kerta Mukti, Medang Asem, Klari, Kutamukti, Kutawaluya, Lemahdhuhur, Loji, Majalaya, Nagasari, Pacing, Pakis , Jaya, Pangkalan, Pedes, Purwasari, Rengasdengklok, Pasirukem, Tanjung Pura, Telaga Sari, Tempuran, Tirta Mulya, Tunggak Jati, Wadas	Cibuaya, Kota Baru, Klawad, Rawa Merta, Sungai Buntu, Tirta Jaya

Tabel 5. Hasil Nilai DBI

Distance Measure	Nilai DBI
Euclidean Distance	0,088
Chebyshev Distance	0,116

Pembahasan

Dataset yang digunakan adalah dataset penyakit ISPA tahun 2019 berisi 50 *record* yang diperoleh dari Dinas Kesehatan Karawang. Dataset ini digunakan untuk membuat model klastering yang diharapkan dapat membantu pemerintah untuk mengetahui daerah penyebaran penyakit ISPA di Kabupaten Karawang dengan metode klastering pada *data mining.* Algoritma yang digunakan pada penelitian ini yaitu K-Medoids dengan dua kali percobaan menggunakan *distance measure* yang berbeda. Adapun distance measure yang digunakan yaitu *Euclidean Distance* dan *Chebyshev Distance*. Penggunaan kedua *distance measure* tersebut bertujuan untuk mendapatkan hasil klastering paling optimal yang akan dievaluasi dengan *Davies Bouldin Index*.

Penelitian dilaksanan mengikuti tahapan pada metodologi *Knowledge Discovery in Database* (KDD) dimana tahapan pertama yang dilakukan yaitu pengumpulan data seperti yang sudah dijelaskan sebelumnya. Kemudian dilakukan seleksi data yang menyisakan 9 dari 25 atribut yang ada pada dataset. Selanjutnya adalah transformasi data dengan metode *Min-max normalization*. Hal ini bertujuan untuk menyetarakan *range* setiap antar atribut dengan skala 0 sampai 1. Hasil transformasi data dapat dilihat pada Tabel 3.

Selanjutnya dilakukan pemodelan dataset menggunakan dengan bantuan *tool* RapidMiner. Hasil anggota klaster masing-masing *distance measure* ditampilkan pada Tabel 4 dimana *Euclidean Distance* menempatkan 6 anggota pada *cluster_0*, 44 anggota pada *cluster_1*, dan 1 anggota pada *cluster_2* sedangkan *Chebyshev Distance* menempatkan 4 anggota pada *cluster_0*, 41 anggota pada *cluster_1*, dan 6 anggota pada *cluster_2*. Perbedaan hasil anggota masing-masing *distance measure* terpaut jauh pada *cluster_2* dimana *Euclidean Distance* hanya menempatkan 1 anggota sedangkan *Chebyshev Distance* menempatkan 6 anggota didalamnya.

Davies Bouldin Index atau DBI digunakan untuk melakukan evaluasi hasil klaster pada masing-masing *distance measure*. Hasil nilai DBI dapat dilihat pada Tabel 5 dimana *Euclidean Distance* mendapatkan nilai DBI sebesar 0,088 sedangkan *Chebyshev Distance* mendapatkan nilai DBI sebesar 0,116. Dari hasil nilai DBI yang didapatkan, dapat dilihat bahwa penggunaan *Euclidean Distance* mendapatkan nilai yang mendekati 0 yang berarti hasil klastering dengan *distance measure* ini memiliki kemiripan yang tinggi dalam satu kelompoknya. Berdasarkan hasil tersebut, dapat dikatakan bahwa penggunaan *Euclidean Distance* pada algoritma K-Medoids lebih optimal jika dibandingkan dengan *Chebyshev Distance.*

Sementara itu, penelitian sebelumnya dengan kasus penentuan status EDGI (*E-Goverment Development Index*) yang dilakukan oleh (Mustofa & Suasana, 2018), menyimpulkan bahwa *Chebyshev Distance* pada K-Medoids berhasil mengoptimalkan penentuan pengelompokkan EDGI jika dibandingkan dengan *Manhattan Distance* dan *Euclidean Distance*. Hal ini membuktikan bahwa hasil optimasi setiap *distance measure* pada algoritma K-Medoids tergantung pada dataset dan permasalahan yang akan diselesaikan. Untuk itu perlu dilakukan uji dataset dengan beberapa *distance measure* sehingga didapatkan hasil terbaik dengan *distance measure* yang tepat.

SIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa *distance measure* yang digunakan sangat berpengaruh terhadap hasil klastering. Penggunaan *Euclidean Distance* pada algoritma K-Medoids menghasilkan klaster optimal jika dibandingkan dengan *Chebyshev Distance*. Dilihat dari nilai *Davies Bouldin Index* yang dihasilkan oleh *Euclidean Distance* sebesar 0,088 yang paling mendekati 0 menunjukkan bahwa hasil klaster dengan *Euclidean Distance* memiliki kemiripan yang tinggi antar objek didalam kelompoknya.

REFERENSI

Bastian, A., Sujadi, H., & Febrianto, G. (2018). Penerapan Algoritma k-means clustering analisis pada penyakit menular manusia. *Analisis Pada Penyakit Menular Manusia*, *14*(1), 28–34.

de la Vega, A., García-Saiz, D., Zorrilla, M., & Sánchez, P. (2020). Lavoisier: A DSL for increasing the level of abstraction of data selection and formatting in data mining. *Journal of Computer Languages*, *60*, 100987.

Ghazal, M. M., & Hammad, A. (2020). Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects. *International Journal of Construction Management*, 1–15. https://doi.org/10.1080/15623599.2020.1738205

Gueorguieva, N., Valova, I., & Georgiev, G. (2017). M&MFCM: fuzzy c-means clustering with mahalanobis and minkowski distance metrics. *Procedia Computer Science*, *114*, 224–233.

Gultom, S., Sriadhi, S., Martiano, M., & Simarmata, J. (2018). Comparison analysis of K-means and K-medoid with Ecluidience distance algorithm, Chanberra distance, and Chebyshev distance for big data clustering. *IOP Conference Series: Materials Science and Engineering*, *420*(1), 12092. IOP Publishing.

Gunawan, I., Anggraeni, G., Rini, E. S., & Mustofa, Y. (2020). Klasterisasi provinsi di Indonesia berbasis perkembangan kasus Covid-19 menggunakan metode K-Medoids. *Seminar Nasional Matematika Dan Pendidikan Matematika (5thSENATIK)*, 301–306. Semarang: Universitas PGRI Semarang Press.

He, L., Agard, B., & Trépanier, M. (2020). A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. *Transportmetrica A: Transport Science*, *16*(1), 56–75.

Irawan, E., Siregar, S. P., Damanik, I. S., & Saragih, I. S. (2020). Implementasi Algoritma K-Medoids untuk Pengelompokkan Sebaran Mahasiswa Baru. *Jurasik (Jurnal Riset Sistem Informasi Dan Teknik Informatika)*, *5*(2), 275–281. https://doi.org/10.30645/jurasik.v5i2.213

Ishak, A., Siregar, K., Ginting, R., & Afif, M. (2020). Orange Software Usage in Data Mining Classification Method on The Dataset Lenses. *IOP Conference Series: Materials Science and Engineering*, *1003*(1), 12113. IOP Publishing.

Juninda, T., Mustasim, & Andri, E. (2019). Penerapan Algoritma K-Medoids untuk Pengelompokan Penyakit di Pekanbaru Riau. *Seminar Nasional Teknologi Informasi, Komunikasi Dan Industri*, *11*(1), 42–49.

Kumar, N., Jain, S., & Chauhan, K. (2019). Knowledge Discovery from Data Mining Techniques. *International Journal of Engineering Research & Technology (IJERT)*, *7*(12), 1–3.

Liu, H., Zhang, X., Zhang, X., & Cui, Y. (2017). Self-adapted mixture distance measure for clustering uncertain data. *Knowledge-Based Systems*, *126*, 33–47.

Miftahuddin, Y., Umaroh, S., & Karim, F. R. (2020). Perbandingan Metode Perhitungan Jarak Euclidean, Haversine, dan Manhattan dalam Penentuan Posisi Karyawan. *Jurnal Tekno Insentif*, *14*(2), 69–77. https://doi.org/10.36787/jti.v14i2.270

Mustofa, Z., & Suasana, I. S. (2018). Algoritma Clustering K-Medoids Pada E-Government Bidang Information And Communication Technology Dalam Penentuan Status EDGI. *Jurnal Teknologi Informasi Dan Komunikasi*, *9*(1), 1–10.

Nahdliyah, M. A., Widiharih, T., & Prahutama, A. (2019). Metode K-Medoids Clustering dengan Validasi Silhouette Index dan C-Index. *JURNAL GAUSSIAN*, *8*(2), 161–170.

Nawrin, S., Rahatur, M., & Akhter, S. (2017). Exploreing K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System. *International Journal of Advanced Computer Science and Applications*, *8*(3), 264–272. https://doi.org/10.14569/ijacsa.2017.080337

Ningrat, D. R., Maruddani, D. A. I., & Wuryandari, T. (2016). Analisis Cluster Dengan Algoritma K-Means Dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi. *None*, *5*(4), 641–650.

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*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

Pandey, A., & Jain, A. (2017). Comparative Analysis of KNN Algorithm using Various Normalization Techniques. *International Journal of Computer Network and Information Security*, *9*(11), 36–42. https://doi.org/10.5815/ijcnis.2017.11.04

Santoso, B., Cholissodin, I., & Setiawan, B. D. (2017). Optimasi K-Means untuk Clustering Kinerja Akademik Dosen Menggunakan Algoritme Genetika. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *1*(12), 1652–1659.

Saputra, D. M., Saputra, D., & Oswari, L. D. (2020). Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method. *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN*, 341–346. Indonesia: Atlantis Press.

Sari, V. R., Firdausi, F., & Azhar, Y. (2020). Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes. *Edumatic: Jurnal Pendidikan Informatika*, *4*(2), 1–9.

Schmidt, C., & Sun, W. N. (2018). Synthesizing agile and knowledge discovery: case study results. *Journal of Computer Information Systems*, *58*(2), 142–150.

Takdirillah, R. (2020). Penerapan Data Mining Menggunakan Algoritma Apriori Terhadap Data Transaksi Penjualan Bisnis Ritel. *Edumatic: Jurnal Pendidikan Informatika*, *4*(1), 37–46.

Tao, X., Wang, R., Chang, R., Li, C., Liu, R., & Zou, J. (2019). Spectral clustering algorithm using density-sensitive distance measure with global and local consistencies. *Knowledge-Based Systems*, *170*, 26–42.

Uska, M., Wirasasmita, R., Usuluddin, U., & Arianti, B. (2020). Evaluation of Rapidminer-Aplication in Data Mining Learning using PeRSIVA Model. *Edumatic: Jurnal Pendidikan Informatika*, *4*(2), 164–171. https://doi.org/10.29408/edumatic.v4i2.2688