ALGORITME *PARTITIONING AROUND MEDOID* (PAM) DENGAN *CALINSKI-HARABASZ INDEX* UNTUK *CLUSTERING* DATA *OUTLIER*

(Partitioning Around Medoid (PAM) Algorithm with Calinski-Harabasz Index for Clustering Data Outlier)

Aliyatussya'ni[1*], Dewi Retno Sari Saputro[2)]

[1,2)]Program Studi Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Sebelas Maret,
Jl. Ir. Sutami No. 36A, Surakarta 57126, Indonesia
Email : aliya.tuss321@student.uns.ac.id, dewiretnoss@staff.uns.ac.id
*)penulis korespondensi

Abstract. The process of gathering information from a mathematical pattern in big data to help make decisions is called data mining. Cluster analysis is a multivariate statistical analysis technique that groups observations based on several variables based on the level of similarity. Clustering is a technique in data mining that aims to group data into several clusters. Data objects that have high similarity will be in the same cluster. Outliers data that is different from other data. In statistics, the presence of this outlier will result in data analysis being biased and not reflecting the actual phenomenon. Partitioning Around Medoid (PAM) or K-Medoid is a non-hierarchical-based clustering algorithm. The steps carried out in the PAM algorithm are grouping the data by dividing the data into n groups. Calinski-Harabasz Index is one of the methods used to determine the best number of clusters. The purpose of this study was to examine the PAM algorithm on data containing outliers and the Calinski-Harabasz Index as a method for selecting the best cluster. The results showed that the PAM algorithm and the Calinski-Harabasz Index have good robustness for outlier data.

Keywords: *Calinski-Harabasz Index, Clustering, Outlier, PAM*

1. Pendahuluan

Data mining merupakan suatu proses yang menggunakan statistik, matematika, kecerdasan buatan dan teknik pembelajaran mesin (*machine learning*) untuk mengekstrak informasi dari database besar [17]. Deskripsi, estimasi, prediksi, klasifikasi, pengklasteran dan asosiasi merupakan bentuk data mining [12]. Pengklasteran atau pengelompokan merupakan sebuah teknik dalam data mining yang bertujuan untuk mengelompokkan objek yang memiliki kesamaan ke dalam kelompok-kelompok (*cluster*). Objek yang memiliki tingkat kemiripan yang tinggi dikelompokkan ke dalam satu kelompok yang sama dan objek yang memiliki tingkat kemiripan yang rendah akan dikelompokkan ke dalam kelompok yang lain. Analisis kelompok atau analisis *cluster* merupakan salah satu teknik analisis *multivariat* yang menganalisis hubungan banyak variabel. *Clustering* memiliki dua pendekatan, yaitu pendekatan *hierarki* dan pendekatan *non hierarki.*

K-Means merupakan salah satu algoritme *clustering* yang berbasis *non hierarki* yang dipublikasi oleh Forgey pada tahun 1965 yang termuat di dalam Kaufmaan dan Rousseeuw [9]. *K-Means* dikenal juga sebagai metode Lloyd-Forgy. Pada tahun 1967, James MacQueen [14] mengembangkan metode *K-Means* sebagai metode sederhana berbasis *centroid*. Algoritme *K-Means* sensitif terhadap *outlier* karena menggunakan nilai rata-rata dalam penentuan pusat *clusternya*. Pada tahun 1990, Kaufmaan dan Rousseeuw [9] melakukan penelitian tentang algoritme *K-Medoid* atau *PAM*. Dalam penelitiannya Kaufmaan dan Rousseeuw [9] menjelaskan bahwa algoritme *PAM* dapat digunakan untuk mengatasi data yang memuat *outlier*. Algoritme *PAM* membutuhkan jumlah *cluster* untuk memudahkan dalam proses pengelompokan. Penentuan jumlah *cluster* dalam algoritme *PAM* dapat ditentukan langsung oleh peneliti, namun hal ini mengakibatkan jumlah *cluster* yang diperoleh kurang akurat, sehingga perlu metode untuk menentukan jumlah *cluster* terbaik.
Pada tahun 2012, Madhulata [15] melakukan penelitian dalam penentuan jumlah *cluster* terbaik dengan kriteria *elbow*. Menurut Aksan *et al.*,[2] kriteria *elbow* mudah dihitung tetapi memberikan hasil yang tidak konsisten. Terdapat beberapa metode untuk menentukan jumlah *cluster* terbaik, salah satunya yaitu *Calinski-Harabasz Index*. Oleh karena itu, pada artikel ini dilakukan kajian tentang *Calinski-Harabasz Index* untuk menghitung jumlah *cluster* dan sensitivitasnya terhadap *outlier*.

2. Metodologi

Penelitian ini merupakan penelitian berbasis teori mengenai pengelompokan data yang memuat *outlier* dengan menerapkan algoritma *PAM* dengan disertai penentuan jumlah *cluster* optimal dengan *Calinski-Harabasz Index*. Studi literatur dilakukan dengan mencari dan mengkaji literatur yang membahas tentang data mining, paper di jurnal dan penelitian yang berhubungan dengan *clustering*, *outlier,* algoritme *PAM,* kriteria *elbow,* dan *Calinski-Harabasz Index*.

3. Hasil dan Pembahasan

Pada penelitian ini akan dibahas mengenai *clustering*, *outlier*, algoritme *PAM,* kriteria *elbow* dan *Calinski-Harabasz Index.*

3.1 *Clustering*

Clustering merupakan suatu proses pengelompokkan hasil *record*, observasi, atau kelas yang memiliki kesamaan atau kemiripan antar objek [13]. *Clustering* merupakan salah satu metode data mining yang bersifat *unsupervised* [1]. Terdapat dua pendekatan *clustering* yaitu pendekatan hierarki dan pendekatan non hierarki atau partisi. Metode pengelompokan hierarki bekerja dengan mengelompokkan objek data berdasarkan

kemiripan terdekat sampai membentuk pohon *cluster* dengan tingkatan objek dari yang paling mirip sampai tidak mirip. Metode pengelompokan hierarki dapat digolongkan menjadi dua jenis yaitu *agglomerative* dan *divisive*. Pengelompokan hierarki *agglomerative* bekerja dengan cara *bottom-up* (penggabungan) sedangkan pengelompokan hierarki *divisive* bekerja dengan cara *top-down* (pemisahan)[7]. Pengelompokan berbasis partisi bekerja dengan memecah dataset menjadi beberapa kelompok, dimana banyaknya *k cluster* sudah ditentukan sebelumnya [8]. *PAM* merupakan salah satu algoritme *clustering* yang berbasis partisi.

3.2 *Outlier*

Outlier atau yang biasa dikenal dengan pencilan merupakan suatu objek pengamatan yang menyimpang atau berbeda dengan objek pengamatan lainnya dalam suatu kumpulan data. Sedangkan menurut Cousineau *et al.* [4] *outlier* adalah pengamatan atau pengukuran yang mencurigakan karena jauh lebih kecil atau jauh lebih besar daripada sebagian besar pengamatan. Pada beberapa kasus, *outlier* sengaja dihilangkan untuk menghindari masalah dalam analisis statistik, akan tetapi pada beberapa kasus lain, adanya *outlier* justru memberikan informasi tertentu sehingga kehadirannya sangat dibutuhkan. Menurut Sihombing *et al.* [16], *outlier* terbagi menjadi dua, yaitu *outlier univariat* atau *outlier multivariat.* Dalam analisis data dengan analisis *multivariat*, deteksi *outlier* dapat dilakukan dengan cara menghitung kuadrat jarak mahalanobis (d_{ij}^2) dari masing-masing pengamatan [18].

3.3 Algoritma *PAM*

Algoritme *PAM* dapat mengatasi masalah yang berhubungan dengan *outlier*, yang cara kerjanya dengan memilih objek suatu titik sebagai perwakilan titik tengah (*centroid*). *Centroid* ini biasa dikenal dengan nama *medoid*. Menurut Defiyanti *et al.* [5] strategi dasar algoritme PAM yaitu untuk menemukan jumlah *cluster k* pada objek *n* dengan terlebih dahulu dan menemukan objek awal (*medoid*) secara acak sebagai perwakilan untuk setiap *cluster*. Berikut diuraikan algoritme *PAM*

1. Menententukan jumlah *cluster k.*
2. Memilih *medoid* (pusat *cluster*) awal secara acak dari objek yang akan di kelompokkan.
3. Menetapkan setiap objek yang tersisa (*non-medoid*) ke dalam *cluster* yang paling dekat dengan *medoid*.
4. Menetapkan objek *non-medoid* berdasarkan jarak terdekat dengan *medoid* dan menghitung total jarak yang diperoleh.
5. Memilih objek *non-medoid* secara acak pada setiap kelompok sebagai kandidat *medoid* baru dan menghitung jarak setiap objek *non-medoid* baru dengan kandidat *medoid* baru.

6. Menetapkan objek berdasarkan jarak terdekat dengan kandidat *medoid* baru dan menghitung total jarak yang diperoleh.
7. Menghitung total selisih simpangan (S). Nilai S dinyatakan dalam persamaan:

$$S = Total \text{ Distance } baru - Total \text{ Distance } Lama$$

   dengan S merupakan simpangan, total *distance* baru merupakan jumlah *cost non-medoid,* dan total *distance* lama merupakan jumlah *cost medoid*. Jika diperoleh $S < 0$, maka kandidat *medoid* baru tersebut berubah menjadi *medoid* baru.

8. Mengulangi langkah (5) sampai dengan langkah (8) hingga tidak terjadi perubahan *medoid*. Iterasi akan berhenti jika diperoleh $S > 0$. Sehingga pada langkah ini diperoleh kelompok beserta anggota kelompoknya masing-masing.

Berdasarkan penelitian yang dilakukan oleh Arora *et al.,* [3] menunjukkan bahwa waktu yang dibutuhkan dan kompleksitas ruang *cluster* algoritme *PAM* jauh lebih baik daripada algoritme *K-Means*. Dataset yang dihasilkan juga menunjukkan bahwa algoritme *PAM* lebih baik dalam semua aspek seperti waktu eksekusi, tidak sensitif terhadap *outlier* dan pengurangan *noise*. Kompleksitas algoritme *PAM* lebih tinggi dibandingkan *dengan K-Means.*

3.4 Kriteria *Elbow*

Kriteria *elbow* merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik [6]. Cara kerja kriteria *elbow* yaitu dengan memilih *cluster* lalu menambahkan nilai *cluster* untuk memperoleh jumlah *cluster* terbaik. Hasil persentase *varians* dari setiap nilai *cluster* ditunjukkan dengan grafik. Menurut Aksan *et al.* [2] kriteria *elbow* memberikan perhitungan yang cepat, tetapi memberikan nilai angka optimal yang tidak konsisten ketika ada tambahan parameter atau fitur.

3.5 *Calinski Harabasz Index*

Penentuan jumlah *cluster* terbaik atau optimum menjadi faktor yang sangat penting pada pengelompokkan menggunakan algoritme *PAM.* Terdapat beberapa pendekatan sederhana yang umum dan bisa digunakan untuk menentukan jumlah *cluster* optimum. Terdapat tiga jenis kriteria dalam penentuan jumlah *cluster* optimum yaitu kriteria eksternal, kriteria internal dan kriteria relatif. Kriteria internal terdiri dari tiga jenis yaitu *intra-cluster similarity*, *inter-cluster similarity* dan *hybrid* (intra dan inter *cluster*). *Davies-Bouldin index* (DBI), *Dunn index*, *Silhouette index*, dan *Calinski-Harabasz Index* termasuk kedalam jenis kriteria internal.

Calinski-Harabasz Index atau yang biasa disingkat *CH Index* adalah indeks evaluasi

berdasarkan derajat dispersi antara *cluster* (Wang *et al.,* [20]). *CH Index* termasuk kedalam jenis kriteria internal *hybrid* dikarenakan merupakan perpaduan antara *intra-cluster similarity, inter-cluster similarity*. Kriteria ini merupakan kriteria penentuan jumlah *cluster* optimum yang menghitung perbandingan antara nilai *Sum of Square Between-cluster* (*SSB*) sebagai *separation* dan nilai *Sum of Square Within-cluster* (*SSW*) sebagai *compactness* yang dikalikan dengan faktor normalisasi. Faktor normalisasi yaitu selisih jumlah data N dengan jumlah klaster *k* dibagi dengan jumlah klaster *k* dikurang satu. Semakin besar nilai *CH* menunjukkan jumlah *cluster* terbaik [10]. *CH* dapat digunakan sebagai metode untuk menentukan jumlah *cluster* terbaik. Langkah langkah menghitung *CH Index* ditulis sebagai

1. Menghitung $\underline{x_l}$ dan $\underline{x}$ dengan persamaan

$$\underline{x_l} = \frac{1}{N_l} \sum_{\underline{x_l} \in C_l} x_i \tag{1}$$

$$\underline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2}$$

dengan *N* menyatakan banyak data, C_l menyatakan *cluster* ke-*l*, $\underline{x_l}$ menyatakan titik pusat *cluster* ke-*l*, x_i menyatakan titik ke-*i* pada cluster ke-*l*, N_l menyatakan jumlah titik pada *cluster* ke-*l*.

2. Menghitung nilai *Sum of Square Beetween-cluster (SSB)* dengan persamaan

$$SSB = \sum_{l=1}^{k} N_l (\underline{x_l} - \underline{x})(\underline{x_l} - \underline{x})^T \tag{3}$$

dengan *N* menyatakan banyak data, *k* menyatakan banyak *cluster*, C_l menyatakan *cluster* ke-*l*, $\underline{x_l}$ menyatakan titik pusat *cluster* ke-*l*, x_i menyatakan titik ke-*i* pada cluster ke-*l*, N_l menyatakan jumlah titik pada *cluster* ke-*l*.

3. Menghitung nilai *Sum of Square Within-cluster (SSW)* dengan persamaan

$$SSW = \sum_{l=1}^{k} \sum_{\underline{x_l} \in C_l} (x_i - c_k)(x_i - \underline{x_l})^T \tag{4}$$

dengan *k* menyatakan banyak *cluster*, C_l menyatakan *cluster* ke-*l*, $\underline{x_l}$ menyatakan titik pusat *cluster* ke-*l*, x_i menyatakan titik ke-*i* pada cluster ke-*l*.

4. Menghitung *CH Index* dengan persamaan

$$CH = \frac{N-k}{k-1} x \frac{SSB}{SSW} \tag{5}$$

dengan *k* menyatakan banyak *cluster*, *N* menyatakan banyak data, *SSW*

menyatakan *Sum of Square Within-cluster*, *SSB* menyatakan *Sum of Square Between-cluster*, *CH* menyatakan *Calinski-Harabasz Index.*

Semakin kecil nilai *SSW* maka semakin dekat hubungan dalam *cluster*, semakin besar nilai *SSB* maka semakin tinggi derajat dispersinya dan semakin besar nilai *CH Index* maka semakin baik efek *clusteringnya* [19]. Pada penelitian yang dilakukan oleh Kingrani *et al.* [11], *CH index* memiliki skor yang tinggi yaitu sebesar 49 untuk mengelompokkan data yang relatif bising (mengandung *outlier*). Masih menurut Kingrani *et al.* [11], *CH Index* juga memiliki skor yang tinggi untuk data yang tidak memuat *outlier* yaitu sebesar 50. Oleh karena itu, *CH Index* merupakan kriteria penentuan jumlah *cluster* optimum yang relatif baik untuk mengatasi data *outlier.*

4. Kesimpulan

Berdasarkan pembahasan dapat disimpulkan bahwa data *outlier* dapat diselesaikan dengan dengan menggunakan algoritme *PAM*. Salah satu metode untuk menentukan jumlah *cluster* yaitu *Calinski-Harabasz Index. Calinski-Harabasz Index* juga memiliki sifat *robust* terhadap *outlier*. Oleh karena itu, algoritme *PAM* dan *Calinski-Harabasz Index* dapat menyelesaikan masalah yang berhubungan dengan *outlier*.

Daftar Pustaka

[1] Agusta, Y., (2007), K-Means – Penerapan, Permasalahan dan Metode Terkait, *Jurnal Sistem dan Informatika*, **3**, 47-60.

[2] Aksan F., Jasi´nski M., Sikorski T., Kaczorowska D., Rezmer J., Suresh V., Leonowicz Z., Kostyła P., Szyma ´nda J., Janik P., (2021), Clustering Methods for Power Quality Measurements in Virtual Power Plant, *Energies 2021*, **14**, 5902. https://doi.org/10.3390/en14185902

[3] Arora, P., Deepali., Varshney, S., (2015), Analysis of K-Means and K-Medoids Algorithm for Big Data. *International Conference on Information Security & Privacy (ICISP2015)*, 11-12 December 2015, Nagpur, INDIA

[4] Cousineau, D., Chartier, S., (2010), Outliers detection and treatment: a review. *International Journal of Psychological Research*, **3(1)**. ISSN impresa (printed) 2011-2084 ISSN electrónica (electronic) 2011-2079

[5] Defiyanti, S., Jajuli, M., Rohmawati, N., (2017), Optimalisasi K-Medoid dalam Pengklasteran Mahasiswa Pelamar Beasiswa dengan Cubic Clustering Criterion. *TEKNOSI*, **3(1)**, 211-218. ISSN 2476 - 8812

[6] Dewi, A.I.C., & Pramita, D.A.K., (2019), Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *JURNAL MATRIX*, **9(3)**.

[7] Han, J., Kamber, M., Pei J., (2012), *Data Mining Concepts and Techniques Third Edition*. San Massachusetts (US): Morgan Kaufmann Publisher.

[8] Johnson, R.A. & Winchern, D. W., (2014), *Applied Multivariate Statistical Analysis.* **6**. London, UK: Pearson.

[9] Kaufman, L. & Rousseauw P. J., (1990), *Finding Groups in Data an Introduction to Cluster Analysis*. Hoboken, New Jersey, Canada: John Wiley & Sons, Inc.

[10] Khairati, A. F., Adlina, A. A., Hertono, G. F., Handari, B. D., (2019), Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA, *PRISMA, Prosiding Seminar Nasional Matematika 2 (2019)*, 161-170. ISSN 2613-9189.

[11] Kingrani, Kumar S., Levene, Mark, Zhang, Dell, (2018), Estimating the number of clusters using diversity, *Artificial Intelligence Research*, **7 (1)**, 15-22. ISSN 1927-6974.

[12] Kusrini., & Luthfi E. T., 2009. *Algoritma Data Mining*. Yogyakarta: Penerbit Andi.

[13] Larose, Daniel, T. and Larose, Chantal D., (2015), *Data Mining and Predictive Analytics. Second Edition*, John Wiley & Sons.

[14] MacQueen, J., (1967),. *Some Methods for Clasification and Analysis of Multivariat Observation*. University of Californa, Los Angeles.

[15] Madhulata, T. S., (2012), An Overview on Clustering Methods. *IOSR Journal of Engineering*, **2(4)**, 719-725.

[16] Sihombing, R. E., Rachmatin, D., Dahlan, J. A., (2019), Program Aplikasi Bahasa R Untuk Pengelompokan Objek Menggunakan Metode K-Medoids Clustering, *Jurnal Eurikamatika*, 7(1), 58-79.

[17] Turban, E., Aronson, J. E., Ting-Peng, l., (2007). *Decision Support Systems and Intelligent Systems Seventh Edition*. New Delhi: Asoke K. Ghosh.

[18] Utami, D., S., & Saputro, D., R., S., (2018). PENGELOMPOKAN DATA YANG MEMUAT PENCILANDENGAN KRITERIA ELBOWDAN KOEFISIEN SILHOUETTE (ALGORITME K-MEDOIDS). *KNPMP III 2018*. ISSN: 2502-6526

[19] Wang, X., Xu, Y., (2019), An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conf. Series: Materials Science and Engineering*, **569,** 052024. doi:10.1088/1757-899X/569/5/052024