

Perbandingan Hybrid Genetic K-Means++ dan Hybrid Genetic K-Medoid Untuk Klasterisasi Dataset Eeg Eyestate

Muhammad Ezar Al Rivan¹; Giovani Prakasa Gandi²; Fendy Novianto Lukman³

^{1,2,3}Teknik Informatika STMIK Global Informatika MDP

¹meedzhar@mdp.ac.id

ABSTRACT

K-Means++ and K-Medoids are data clustering methods. The data cluster speed is determined by the iteration value, the lower the iteration value, the faster the data clustering is done. Data clustering performance can be optimized to get more optimal clustering results. One algorithm that can optimize cluster speed is Genetic Algorithm (GA). The dataset used in the study is a dataset of EEG Eyestate. The optimization results before hybrid GA on K-Means++ are the iteration average values is 11.6 to 5,15, and in K-Medoid are the iteration average values decreased from 5.9 to 5.2. Based on the comparison of GA K-Means++ and GA K-Medoids iterations, it can be concluded that GA - K-Means++ better.

Keywords: EEG Eyestate Dataset, GA, K-Means++, K-Medoids, Cluster

ABSTRAK

K-Means++ dan K-Medoid adalah metode pengklasteran data. Tingkat kecepatan kluster data ditentukan oleh nilai iterasi, semakin rendah nilai iterasi maka semakin cepat pengklasteran data yang dilakukan. Kinerja pengklasteran data dapat dioptimasi untuk mendapatkan hasil pengklasteran yang lebih optimal. Salah satu algoritma yang mampu mengoptimasi kecepatan kluster adalah Genetic Algorithm (GA). Dataset yang digunakan pada penelitian adalah dataset EEG Eyestate. Hasil optimasi sebelum dan sesudah hybrid GA pada K-Means++ terlihat pada nilai rata-rata iterasi dari 11,6 menjadi 5,15 dan hasil optimasi K-Medoids sebelum dan sesudah hybrid GA terlihat pada nilai rata-rata iterasi dari 5,9 menjadi 5,2. Berdasarkan hasil perbandingan nilai rata-rata iterasi GA K-Means++ dan GA K-Medoids maka dapat disimpulkan bahwa GA - K-Means++ lebih baik

Kata kunci: Dataset EEG Eyestate, GA, K-Means++, K-Medoids, Kluster

1. PENDAHULUAN

Metode untuk pengelompokan dengan tujuan mengelompokan objek sudah digunakan di berbagai bidang ilmu pengetahuan. Dalam mengelompokan data harus memiliki informasi dari data tersebut untuk diklasifikasi data yang diolah. Jika data yang ada tidak diketahui cara untuk mengolahnya, maka metode klasifikasi tidak dapat digunakan, sehingga membuat metode pengklasteran lebih tepat dipakai untuk menyelesaikan masalah yang ada.

Clustering dengan menggunakan kombinasi algoritma *clustering* dan *genetic* dilakukan oleh [1] dan [2] menggunakan *Fuzzy C-Means* dan *Genetic Algorithm*. Selain *Fuzzy C-Means* ada beberapa penelitian yang digunakan untuk klasterisasi menggunakan algoritma *K-Means++* dan *K-Medoid*. Penelitian yang dilakukan oleh [3] mengenai “*A New Hybrid Clustering Technique Based On Mini-batch K-means And K-means++ For Analysing Big Data*” didapatkan hasil bahwa algoritma *K-Means++* bisa mendapatkan *centroid* yang bagus dan algoritma *clustering hybrid* ini mengurangi komputasi waktu. Penelitian lain dilakukan oleh [4] mengenai “*A Genetic K-Medoids Clustering Algorithm*” didapatkan hasil bahwa algoritma yang diusulkan dapat secara efisien mengembangkan partisi yang tepat sementara tidak membuat asumsi prioritas tentang jumlah *cluster* yang ada di dataset [5]. Selain itu perbandingan *K-Means* dan *K-Means++* dilakukan oleh [6]. Pada penelitian ini menunjukkan bahwa *K-Means++* mengungguli *K-Means*. Pada penelitian [7] *K-Means++* digunakan untuk *cluster* pasien dengan gejala penyakit.

Selain *K-Means++* terdapat algoritma *cluster* yaitu *K-Medoid*. Algoritma ini digunakan pada penelitian [8]. Pada penelitian tersebut *K-Medoid* digunakan untuk *cluster* pasien kanker berdasarkan bentuk protein. Pada penelitian [9], *K-Medoid* digunakan untuk *cluster* strategi pemasaran. Pada penelitian [10] *K-Medoid* dikombinasikan dengan menggunakan *heuristic*.

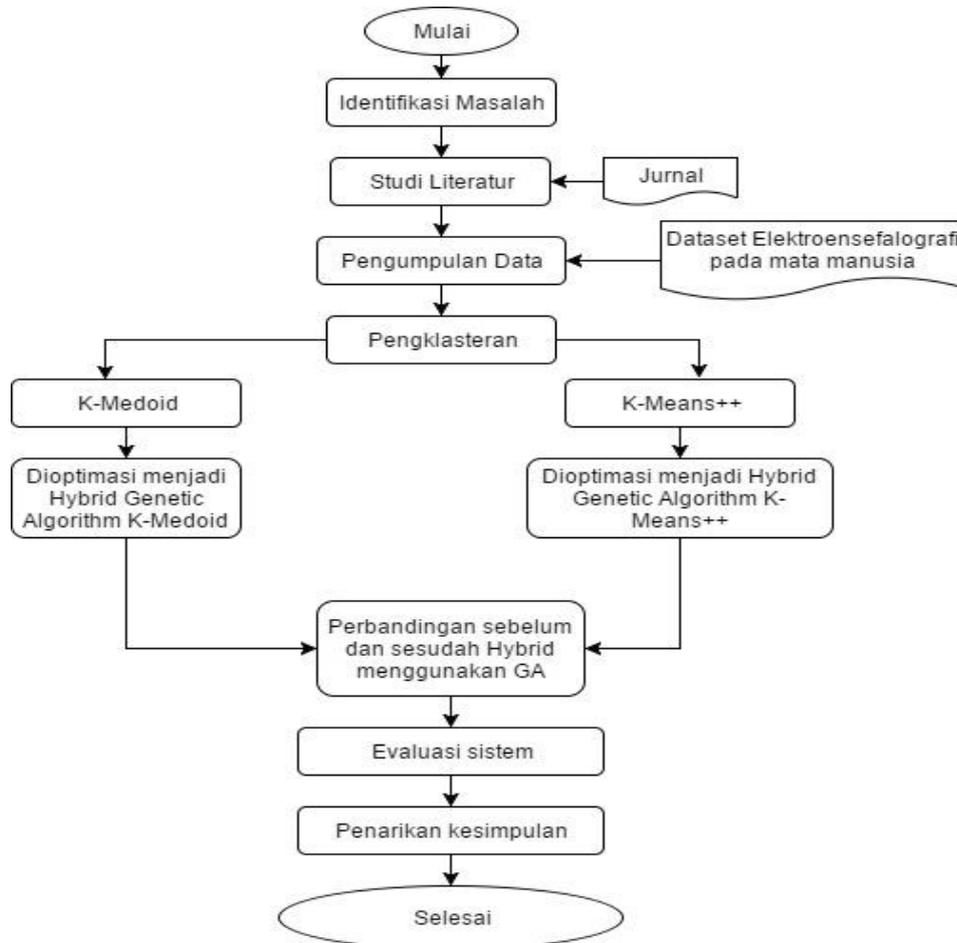
Penggabungan Algoritma Genetika dan *Fuzzy C-Means* dilakukan oleh [11] untuk penentuan kompetisi mahasiswa berhasil memperkecil total *cost* dan jumlah variabel. Presentase rata-rata kesesuaian pada pengujian yang dilakukan dengan menggunakan data kuesioner terhadap mahasiswa adalah sebesar 88.89%. Pada penelitian [12] FCM digunakan untuk *cluster* nilai akhir beserta dengan GA. Pada penelitian yang dilakukan oleh [13] dalam membandingkan iterasi FCM, GA FCM, *K-Means*, GA *K-Means* didapat rata-rata iterasi FCM sebesar 17,1, GA FCM sebesar 6,65, *K-Means* sebesar 10,85, dan GA *K-Means* sebesar 7,35. Penelitian terkait dengan perbandingan *K-Means* dan *K-Medoid* dilakukan oleh [14].

Seluruh *dataset* yang akan digunakan didapatkan dari donasi oleh Oliver Roesler [5]. *Dataset* terdiri atas 7 *corpus* dimana masing-masing *corpus* merupakan hasil perekaman selama kurang lebih 2 menit dari 4 orang yang berbeda. Masing-masing orang menghasilkan 2 *corpus*. Terdapat kamera yang memeriksa keadaan mata dari pelaku yang kemudian akan ditambahkan ke dalam *dataset* sebagai target. *Dataset corpus* memiliki 15 *feature*, 14 atribut nilai *real* EEG dari *channel* AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4 dan 1 nilai *integer* target keadaan mata. 0 menandakan keadaan mata terbuka dan 1 menandakan keadaan mata tertutup. Seluruh keadaan saat mata belum sepenuhnya tertutup dikategorikan sebagai mata terbuka [15]

Berdasarkan uraian tersebut, maka penulis mengangkat topik penelitian mengenai metode *Hybrid Genetic Algorithm K-Means++* dan *Hybrid Genetic Algorithm K-Medoid* untuk pengklasteran dataset EEG pada mata manusia. Adapun hasil yang didapat dalam penelitian ini adalah hasil pengklasteran untuk melihat perbedaan kecepatan dalam pengklasteran sebelum dan sesudah dioptimasi menggunakan algoritma genetika.

2. METODE/PERANCANGAN PENELITIAN

Pada penelitian ini dilakukan dengan menggunakan beberapa tahapan yang harus dilakukan secara berurutan guna mencapai hasil yang maksimal seperti studi literatur, pengumpulan data, pengklasteran dataset menggunakan metode *K-Means++* dan *K-Medoid* sebelum dan sesudah dioptimasi, dan evaluasi sistem. Pada tahapan perancangan sistem dalam pengklasteran *dataset* menggunakan metode *K-Means++* dan *K-Medoid* dapat dilihat pada Gambar 1.



Gambar 1. Metodologi Penelitian

Metodologi yang digunakan penulis untuk melakukan penelitian adalah sebagai berikut:

1. Identifikasi Masalah
Pada tahap ini, dimulai dengan mengidentifikasi masalah untuk memahami masalah yang diangkat ke dalam penelitian dan solusi untuk memecahkan masalah tersebut.
2. Studi Literatur
Pada tahap ini peneliti akan melakukan studi literatur dengan mengumpulkan informasi berasal dari jurnal – jurnal penelitian sebelumnya yang berkaitan dengan masalah penelitian.
3. Pengumpulan Data
Dalam tahap ini peneliti akan mengumpulkan data berupa dataset *numeric* berjumlah 14.980 dan 15 atribut yang merupakan posisi sensor *neuroheadset Emotiv EEG*
4. Pengklasteran

Pada tahapan ini, dilakukan pengklasteran menggunakan 2 metode *K-Means++* dan *K-Medoid*, lalu kedua metode tersebut dioptimasi dengan GA.

5. Evaluasi sistem

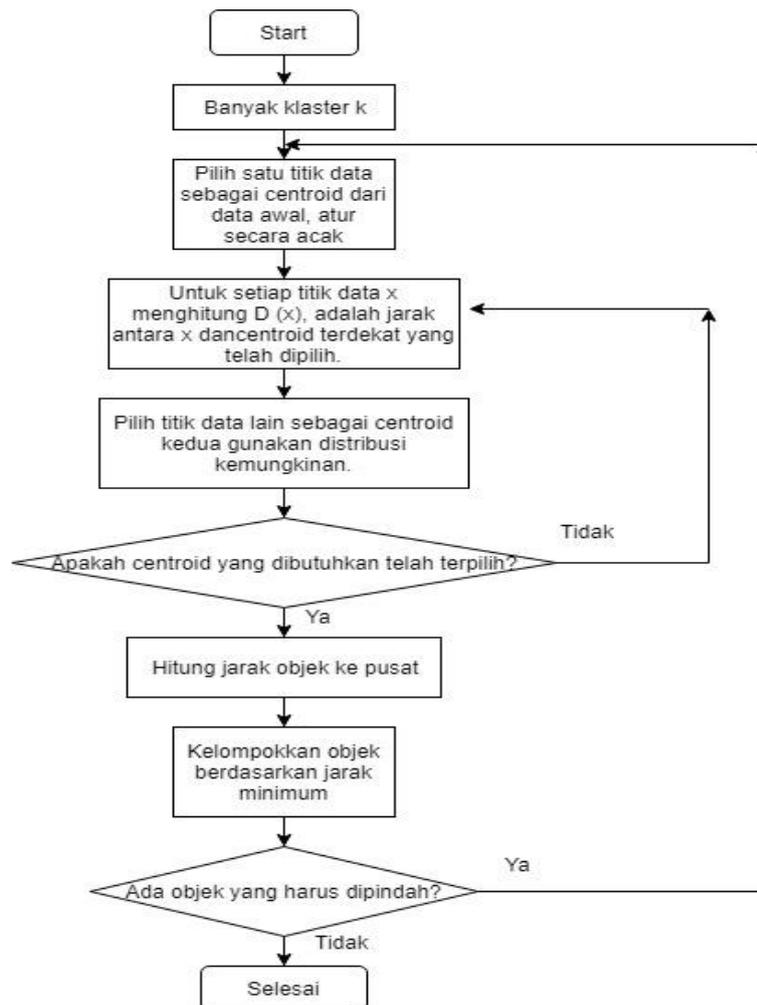
Pada tahap ini dilakukan pengujian terhadap perbandingan klusterisasi metode *K-Means++* dan *K-Medoid* ketika sebelum dan sesudah dioptimasi menggunakan *Hybrid Genetic Algorithm* untuk mengukur iterasi yang kemudian dicatat untuk dijadikan hasil pengujian.

2.1. Data

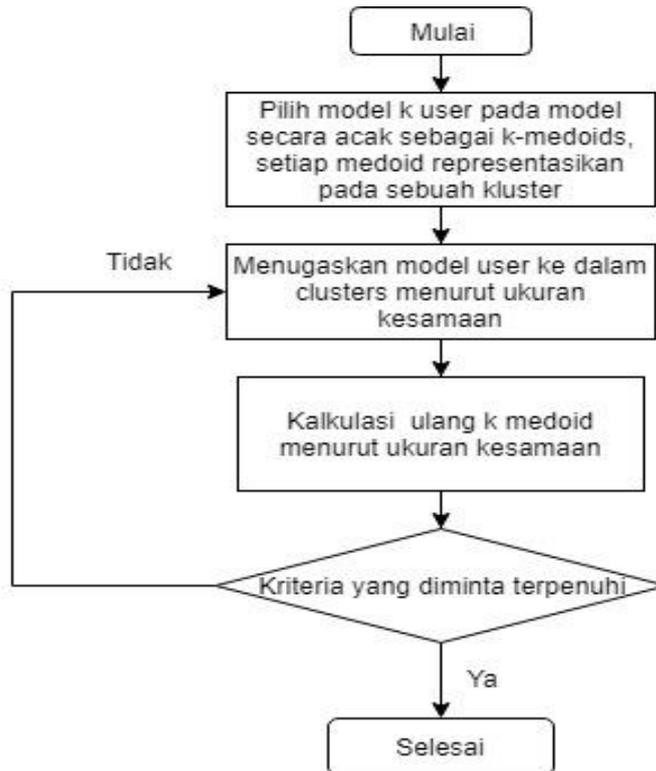
Pada penelitian ini dilakukan pengumpulan data yang berupa dataset *numeric* berjumlah 14.980 dan 15 atribut yang merupakan posisi sensor *neuroheadset Emotiv EEG* yang didapat pada laman resmi *UCI Machine Learning Repository* (2014). Namun, saat tahap klusterisasi, *dataset* yang digunakan berjumlah 14980 dan hanya mengambil 14 atribut awal untuk didapatkan hasil pengujian berupa jumlah iterasi yang selanjutnya dioptimasi menggunakan GA.

2.2. Perancangan

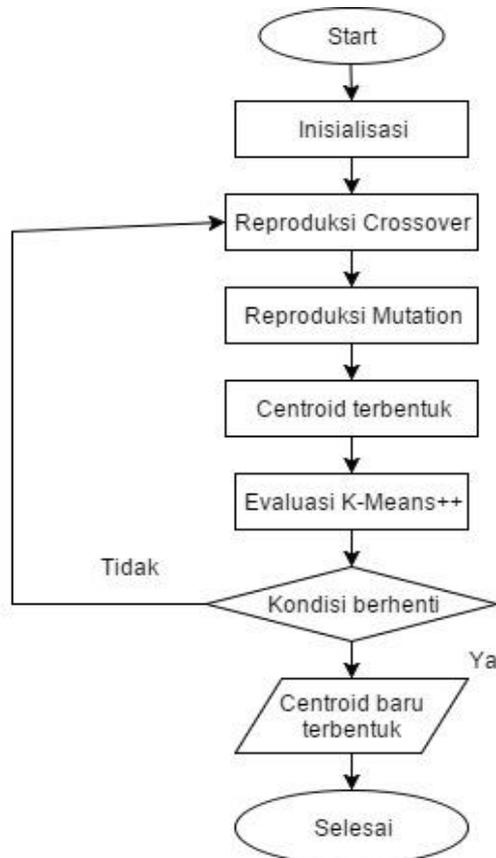
Pada tahap ini penulis merancang alur kerja untuk tahap klusterisasi dataset menggunakan metode *K-Means++* pada gambar 2, *K-Medoid* pada gambar 3, *K-Means++* dioptimasi menggunakan GA pada gambar 4 dan *K-Medoid* sesudah dioptimasi menggunakan GA pada gambar 5.



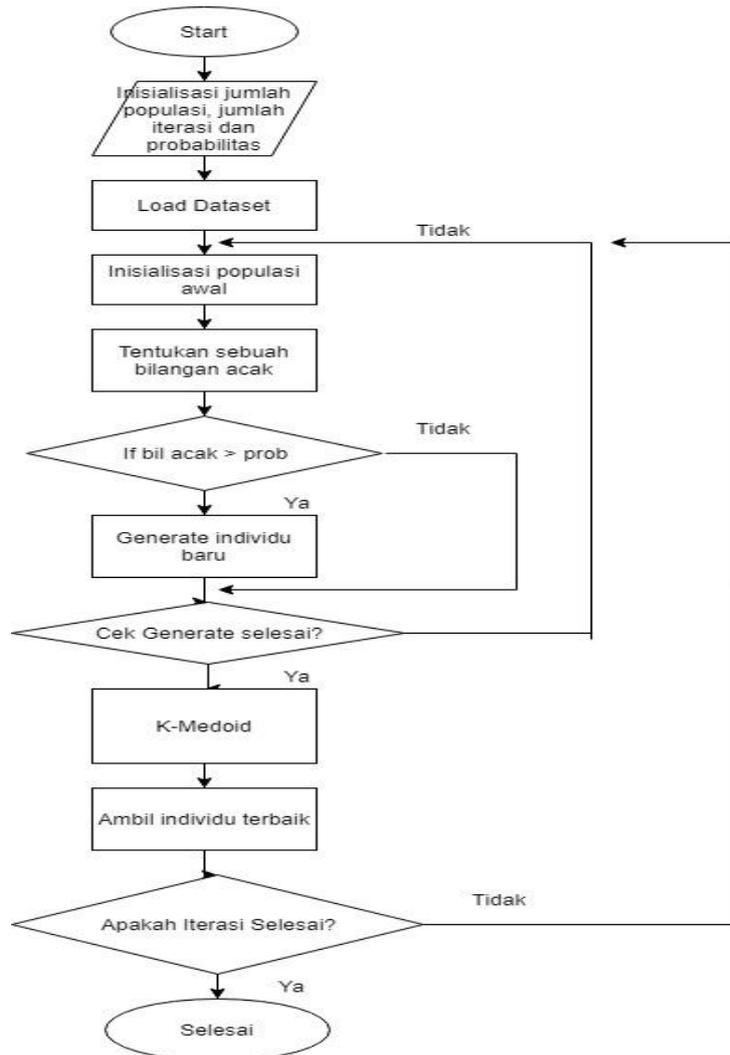
Gambar 2. Diagram Alir K-Means++



Gambar 3. Diagram Alir K-Medoid



Gambar 4. Diagram Alir K-Means++ dengan Optimasi Menggunakan Algoritma Genetika



Gambar 5. Diagram Alir K-Medoid Dioptimasi Menggunakan Algoritma Genetika

3. HASIL DAN PEMBAHASAN

Pada tahap ini dilakukan klusterisasi menggunakan metode *K-Means++* dan *K-Medoid* dimana hasil yang didapatkan berupa perbandingan iterasi lalu dilakukan optimasi menggunakan GA. Pada penelitian ini, pengujian dilakukan sebanyak 20 kali dengan iterasi maksimum sebanyak 100 kali dan jumlah kluster 2 pada metode *K-Means++* menggunakan 3 varian *distance* yaitu *Cityblock*, *Sqeclidean* dan *Cosine* dan metode *K-Medoid*. Untuk hasil pengujian metode *K-Means++* dengan varian *distance Cityblock*, *Sqeclidean* dan *Cosine* dapat dilihat pada Tabel 1.

Tabel 1. Hasil pengujian metode *K-Means++* (Varian *distance Cityblock*, *Sqeclidean*, dan *Cosine*) dan *K-Medoid*

	Iterasi <i>K-Means++</i>			Iterasi <i>K-Medoid</i>
	<i>cityblock</i>	<i>sqeclidean</i>	<i>cosine</i>	
Rata-rata Iterasi	11,6	1,9	2.85	5,9
Iterasi Terbesar	25	11	15	8
Iterasi Terkecil	1	1	1	4

Jumlah iterasi dan rata-rata iterasi untuk GA - *K-Means++* menggunakan *Cityblock Distance* dengan Pc 0,2 dan jumlah individu terdiri dari 10, 20, 30, 40, dan 50 dapat dilihat pada Tabel 2.

Tabel 2. Perbandingan Iterasi GA – *K-Means++ cityblock distance* dengan Pc = 0,2

	Iterasi Maksimum = 100 dan jumlah kluster = 2				
	Jumlah Individu = 10	Jumlah Individu = 20	Jumlah Individu = 30	Jumlah Individu = 40	Jumlah Individu = 50
Rata-rata Iterasi	5,15	5,95	5,95	6,25	6,75
Iterasi Terbesar	7	8	8	8	9
Iterasi Terkecil	4	4	4	4	4

Jumlah iterasi dan rata-rata iterasi untuk GA - *K-Means++* menggunakan *Cityblock Distance* dengan Pc 0,4 dan jumlah individu terdiri dari 10, 20, 30, 40, dan 50 dapat dilihat pada Tabel 3.

Tabel 3. Perbandingan Iterasi GA – *K-Means++ cityblock distance* dengan Pc = 0,4

	Iterasi Maksimum = 100 dan jumlah kluster = 2				
	Jumlah Individu = 10	Jumlah Individu = 20	Jumlah Individu = 30	Jumlah Individu = 40	Jumlah Individu = 50
Rata-rata Iterasi	5,65	5,45	5,8	5,95	6,9
Iterasi Terbesar	7	7	8	8	9
Iterasi Terkecil	4	4	4	4	4

Jumlah iterasi dan rata-rata iterasi untuk GA *K-Means++* menggunakan *Cityblock Distance* dengan Pc 0,6 dan jumlah individu terdiri dari 10, 20, 30, 40, dan 50 dapat dilihat pada Tabel 4.

Tabel 4. Perbandingan Iterasi GA – *K-Means++ cityblock distance* dengan Pc = 0,6

	Iterasi Maksimum = 100 dan jumlah kluster = 2				
	Jumlah Individu = 10	Jumlah Individu = 20	Jumlah Individu = 30	Jumlah Individu = 40	Jumlah Individu = 50
Rata-rata Iterasi	5,3	5,8	5,9	5,95	7,15
Iterasi Terbesar	7	8	7	9	9
Iterasi Terkecil	4	4	4	4	5

Jumlah iterasi dan rata-rata iterasi untuk GA *K-Means++* menggunakan *Cityblock Distance* dengan Pc 0,8 dan jumlah individu terdiri dari 10, 20, 30, 40, dan 50 dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan Iterasi GA – *K-Means++ cityblock distance* dengan $P_c = 0,8$

	Iterasi Maksimum = 100 dan jumlah kluster = 2				
	Jumlah Individu = 10	Jumlah Individu = 20	Jumlah Individu = 30	Jumlah Individu = 40	Jumlah Individu = 50
Rata-rata Iterasi	5,55	5,95	5,7	5,95	7,15
Iterasi Terbesar	7	8	7	8	9
Iterasi Terkecil	4	4	4	4	4

Jumlah iterasi dan rata-rata iterasi untuk GA *K-Medoids* dengan $P_c 0,2$ dan jumlah individu terdiri dari 10, 20, 30, 40, dan 50 dapat dilihat pada Tabel 6.

Tabel 6. Perbandingan Iterasi GA – *K-Medoid* dengan $P_c = 0,2$

	Iterasi Maksimum = 100 dan jumlah kluster = 2				
	Jumlah Individu = 10	Jumlah Individu = 20	Jumlah Individu = 30	Jumlah Individu = 40	Jumlah Individu = 50
Rata-rata Iterasi	5,25	5,9	5,9	5,75	7,1
Iterasi Terbesar	7	8	7	8	9
Iterasi Terkecil	4	4	4	4	5

Jumlah iterasi dan rata-rata iterasi untuk GA *K-Medoids* dengan $P_c 0,4$ dan jumlah individu terdiri dari 10, 20, 30, 40, dan 50 dapat dilihat pada Tabel 7.

Tabel 7. Perbandingan Iterasi GA – *K-Medoid* dengan $P_c = 0,4$

	Iterasi Maksimum = 100 dan jumlah kluster = 2				
	Jumlah Individu = 10	Jumlah Individu = 20	Jumlah Individu = 30	Jumlah Individu = 40	Jumlah Individu = 50
Rata-rata Iterasi	5,55	5,95	5,7	5,95	7,15
Iterasi Terbesar	7	8	7	8	9
Iterasi Terkecil	4	4	4	4	4

Jumlah iterasi dan rata-rata iterasi untuk GA *K-Medoids* dengan $P_c 0,6$ dan jumlah individu terdiri dari 10, 20, 30, 40, dan 50 dapat dilihat pada Tabel 8.

Tabel 8. Perbandingan Iterasi GA – *K-Medoid* dengan $P_c = 0,6$

	Iterasi Maksimum = 100 dan jumlah kluster = 2				
	Jumlah Individu = 10	Jumlah Individu = 20	Jumlah Individu = 30	Jumlah Individu = 40	Jumlah Individu = 50
Rata-rata Iterasi	5,2	5,85	5,95	6,45	6,8
Iterasi Terbesar	6	8	7	9	9
Iterasi Terkecil	4	4	4	4	4

Jumlah iterasi dan rata-rata iterasi untuk GA - *K-Medoids* dengan $P_c 0,8$ dan jumlah individu terdiri dari 10, 20, 30, 40, dan 50 dapat dilihat pada Tabel 9.

Tabel 9. Perbandingan Iterasi GA – *K-Medoid* dengan $P_c = 0,8$

	Iterasi Maksimum = 100 dan jumlah kluster = 2				
	Jumlah Individu = 10	Jumlah Individu = 20	Jumlah Individu = 30	Jumlah Individu = 40	Jumlah Individu = 50
Rata-rata Iterasi	5,65	5,45	5,8	5,95	6,9
Iterasi Terbesar	7	8	8	8	9
Iterasi Terkecil	4	4	4	4	4

Perbandingan jumlah iterasi algoritma *K-Means++* sebelum dan sesudah optimasi dapat dilihat pada Tabel 10.

Tabel 10. Perbandingan Iterasi *K-Means++* sebelum dan sesudah optimasi

	<i>K-Means++</i>	<i>K-Means++ dan GA</i>
Rata-rata Iterasi	11,6	5,15
Iterasi Terbesar	25	7
Iterasi Terkecil	1	4

Perbandingan jumlah iterasi algoritma *K-Medoid* sebelum dan sesudah optimasi dapat dilihat pada Tabel 11.

Tabel 11. Perbandingan Iterasi *K-Medoid* sebelum dan sesudah optimasi

	<i>K-Medoid</i>	<i>K-Medoid dan GA</i>
Rata-rata Iterasi	5,9	5,2
Iterasi Terbesar	8	9
Iterasi Terkecil	4	4

4. KESIMPULAN DAN SARAN

Bedasarkan hasil penelitian ini dapat disimpulkan bahwa GA dapat mengurangi komputasi pada *K-Means++* dan *K-Medoid*. Pengurangan terbesar terjadi pada *K-Means++* - GA dari rata-rata 11,6 menjadi 5,15 sedangkan pada *K-Medoid* - GA pengurangan terjadi dari 5,9 ke 5,2. Secara keseluruhan algoritma *K-Means++* - GA merupakan algoritma yang terbaik untuk klasterisasi dataset EEG Eye State.

DAFTAR PUSTAKA

- [1] Yong-Guo Liu, Ke-Fei Chen, and Xue-Ming Li, "A hybrid genetic based clustering algorithm," Proc. Third Int. Conf. Mach. Learning Cybern., no. August, pp. 1677–1682, 2005, doi: 10.1109/icmlc.2004.1382045.
- [2] A. Al Malki, M. M. Rizk, M. A. El-Shorbagy, and A. A. Mousa, "Hybrid Genetic Algorithm with K-Means for Clustering Problems," Open J. Optim., vol. 05, no. 02, pp. 71–83, 2016, doi: 10.4236/ojop.2016.52009.
- [3] G. Yamini and B. R. Devi, "A New Hybrid Clustering Technique Based On Mini-batch K-means And K-means++ For Analysing Big Data.," Int. J. Recent Res. Asp., no. April, pp. 203–208, 2018.
- [4] W. Sheng and X. Liu, "A Genetic K-medoids Clustering Algorithm," J. Heuristics, vol. 12, no. 6, pp. 447–466, 2006.
- [5] O. Roesler, "EEG Eye State Data Set," Baden-Wuerttemberg Cooperative State University (DHBW), 2013. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State>.
- [6] B. Aubaidan, M. Mohd, M. Albared, and F. Author, "Comparative study of k-means and k-means++ clustering algorithms on crime domain," J. Comput. Sci., vol. 10, no. 7, pp. 1197–1206, 2014, doi: 10.3844/jcssp.2014.1197.1206.
- [7] C. M. Fikri, F. Eka, M. Agustin, and F. Mintarsih, "Pegawai Menggunakan Algoritma K-Means ++ Dan Cop-Kmeans Untuk Merencanakan Program Pemeliharaan Kesehatan Pegawai," J. Pesudocode, vol. IV, pp. 9–17, 2017.
- [8] L. P. Rizby, Marji, and L. Muflikhah, "Clustering Pasien Kanker Berdasarkan Struktur Protein Dalam Tubuh," J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 2, no. 10, pp. 3810–3816, 2018.
- [9] W. A. Triyanto, "Algoritma K-Medoids Untuk Penentuan Strategi Pemasaran Produk," Simetris J. Tek. Mesin, Elektro dan Ilmu Komput., vol. 6, no. 1, p. 183, 2015, doi: 10.24176/simet.v6i1.254.
- [10] M. Nascimento, F. Toledo, and A. Carvalho, "A hybrid heuristic for the k-medoids clustering problem," GECCO'12 - Proc. 14th Int. Conf. Genet. Evol. Comput., pp. 417–424, 2012, doi: 10.1145/2330163.2330223.
- [11] R. Hadi, I. K. Gede Darma Putra, and I. N. Satya Kumara, "Penentuan Kompetensi Mahasiswa dengan Algoritma Genetik dan Metode Fuzzy C-Means," Maj. Ilm. Teknol. Elektro, vol. 15, no. 2, pp. 101–106, 2017, doi: 10.24843/mite.1502.15.
- [12] P. E. Mas'udia and R. Wardoyo, "Optimasi Cluster Pada Fuzzy C-Means Menggunakan Algoritma Genetika Untuk Menentukan Nilai Akhir," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 6, no. 1, pp. 101–110, 2013, doi: 10.22146/ijccs.2145.
- [13] M. E. Al Rivan, S. Steven, and W. Tanzil, "Optimasi Fuzzy C-Means dan K-Means Menggunakan Algoritma Genetika untuk Pengklasteran Dataset Diabetic Retinopathy," J. Teknol. Inf. dan Ilmu Komput., 2020, doi: 10.25126/jtiik.2020711872.

- [14] V. A. P. S, “Perbandingan Algoritma K-Means dan Algoritma K-Medoids dalam Pengelompokan Komoditas Peternakan di Provinsi Jawa Tengah Tahun 2015,” Tugas Akhir Jur. Stat. Fak. Mat. dan Ilmu Pengetah. Alam Univ. Islam Inndonesia Yogyakarta, vol. 53, no. 9, pp. 1689–1699, 2018, doi: 10.1017/CBO9781107415324.004.
- [15] E. C. Prakoso, U. N. Wisesty, and . J., “Klasifikasi Keadaan Mata Berdasarkan sinyal EEG menggunakan Extreme Learning Machines,” *Indones. J. Comput.*, vol. 1, no. 2, p. 97, 2016, doi: 10.21108/indojc.2016.1.2.105.