ALGORITMA PAM (*Partitioning Around Medoids*)-*LITE* DENGAN PROGRAM R UNTUK DATA BERUKURAN BESAR

(Tesis)

Oleh

RIZKI AGUNG WIBOWO
NPM 2027031005

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

ABSTRAK

Algoritma PAM (*Partitioning Around Medoids*)-*Lite* dengan Program R untuk Data Berukuran Besar

Oleh

Rizki Agung Wibowo

Analisis klaster k-medoid atau dikenal sebagai PAM (*Partitioning Around Medoid*) adalah salah satu analisis klaster berbasis partisi. PAM memiliki keunggulan yaitu *robust* terhadap data pencilan akan tetapi memiliki kelemahan yaitu tingkat komputasi yang buruk pada data berukuran besar, sehingga dikembangkan algoritma PAM-*Lite* untuk menjadi alternatif PAM. Pada penelitian ini dilakukan konstruksi *function* PAM-*Lite* pada program R karena belum ditemukan *package* yang didalamnya terdapat *function* PAM-*Lite* kemudian menguji keefektifannya menggunakan data simulasi. Berdasarkan hasil yang didapat, *function* PAM-*Lite* yang telah dikonstruksi dapat diproses dengan baik pada program R dan berdasarkan R-*square*, lebar *silhouette* dan waktu proses, PAM-*Lite* lebih efisien digunakan jika dibandingkan dengan PAM pada data berukuran besar.

Kata kunci : Data pencilan, Analisis klaster, PAM-*Lite*, Pemrograman R

ABSTRACT

PAM (*Partitioning Around Medoids*)-*Lite* algorithm with R for a Massive Data

By

Rizki Agung Wibowo

K-Medoid cluster analysis or known as PAM (Partitioning Around Medoid) is a partition-based cluster analysis method. This method has the advantage namely robust against outliers, however it also has disadvantage i.e. it is worse in computation when applied on large data, for this reason the PAM-Lite algorithm was developed to be an alternative for clustering large data. In this study, we constructed an R function (i.e. a block of codes which only runs when it is called) of PAM-Lite algorithm considering the unavailability of R packages that contains PAM-Lite function. The effectiveness of the constructed function then was tested using basic cluster benchmark data. Based on the result, PAM-Lite function can process properly in R program and is more efficient than PAM based on the R Square, Silhouette width and processing time on large data.

Keyword : Outlier, Cluster Analysis, PAM-Lite, R Programing

ALGORITMA PAM (*Partitioning Around Medoids*)-*LITE* DENGAN PROGRAM R UNTUK DATA BERUKURAN BESAR

Oleh

Rizki Agung Wibowo

Tesis

Sebagai Salah Satu Syarat untuk Mencapai Gelar
MAGISTER MATEMATIKA

Pada
Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

PROGRAM STUDI MAGISTER MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

Judul Tesis	: **ALGORITMA PAM (*Partitioning Around Medoids*)-*LITE* DENGAN PROGRAM R UNTUK DATA BERUKURAN BESAR**
Nama Mahasiswa	: Rizki Agung Wibowo
Nomor Pokok Mahasiswa	: 2027031005
Program Studi	: Magister Matematika
Jurusan	: Matematika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Dr. Khoirin Nisa, S.Si., M.Si.	**Ir. Warsono, M.S., Ph.D.**
NIP 19740726 200003 2 001	NIP 19630216 198703 1 003

2. Ketua Program Studi Magister Matematika

Dr. Asmiati, S.Si., M.Si.
NIP 19760411 200012 2 001

MENGESAHKAN

1. Tim Penguji

Ketua : **Dr. Khoirin Nisa, S.Si., M.Si.** .....

Sekretaris : **Ir. Warsono, M.S., Ph.D.** .....

Penguji Anggota : **1. Prof. Drs. Mustofa Usman, M.A., Ph.D.** .....

2. Prof. Dr. La Zakaria, S.Si., M.Sc. .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T.
NIP 19740705 200003 1 001

3. Direktur Program Pascasarjana

Prof. Dr. Ir. Ahmad Saudi Samosir, S.T., M.T.
NIP 19710413 199803 1 005

Tanggal Lulus Ujian Tesis : **01 Juli 2022**

PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama	: **Rizki Agung Wibowo**
Nomor Pokok Mahasiswa	: **2027031005**
Program Studi	: Magister Matematika
Jurusan	: Matematika

Dengan ini menyatakan bahwa tesis saya yang berjudul "**ALGORITMA PAM (*Partitioning Around Medoids*)-*LITE* DENGAN PROGRAM R UNTUK DATA BERUKURAN BESAR**" adalah hasil pekerjaan saya sendiri. Semua hasil tulisan dalam tesis ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila kemudian hari terbukti bahwa tesis ini merupakan hasil salinan atau telah dibuat orang lain, maka saya bersedia menerima sanksi sesuai ketentuan akademik yang berlaku.

Bandar Lampung, 01 Juli 2022
Penulis,

Rizki Agung Wibowo
NPM. 2027031005

RIWAYAT HIDUP

Penulis dilahirkan di Pringsewu pada tanggal 23 Mei 1998, sebagai anak pertama dari pasangan Bapak Priyo Satmono dan Ibu Sri Sukarni serta kakak dari Sekar Arum Purbo Kinasih.

Penulis telah menempuh pendidikan di Taman Kanak-kanak (TK) Al-Azhar 7 Hajimena pada tahun 2003-2004, Sekolah Dasar Negeri 1 Rajabasa Raya pada tahun 2004-2010, Sekolah Menengah Pertama Negeri (SMPN) 8 Bandar Lampung pada tahun 2010-2013, dan Sekolah Menengah Atas Negeri (SMAN) 14 Bandar Lampung pada tahun 2013-2016.

Pada tahun 2016 penulis terdaftar sebagai Mahasiswa Program Studi S1 Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung melalui jalur SNMPTN dan lulus sebagai sarjana matematika pada tahun 2020. Pada tahun 2020 penulis berkesempatan untuk melanjutkan pendidikan di program studi Magister Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung, selama menjadi mahasiswa Magister penulis cukup aktif mengikuti seminar nasional dan internasional, juga berkesempatan untuk menulis buku bersama Ibu Dr. Khoirin Nisa, S.Si., M.Si. yang berjudul “Simulasi Data Statistik Menggunakan R”

KATA MUTIARA

"Hai orang-orang beriman, jadikanlah sabar dan shalat sebagai penolongmu, sesungguhnya Allah beserta orang-orang yang sabar"
(Qs. Al-Baqarah: 153)

"Perumpamaan (nafkah yang dikeluarkan oleh) orang-orang yang menafkahkan hartanya di jalan Allah adalah serupa dengan sebutir benih yang menumbuhkan tujuh bulir, pada tiap-tiap bulir seratus biji. Allah melipat gandakan (ganjaran) bagi siapa yang Dia kehendaki. Dan Allah Maha Luas (karunia-Nya) lagi Maha Mengetahui."
(Qs. Al-Baqarah: 261)

PERSEMBAHAN

Dengan mengucap puji dan syukur kehadirat Allah SWT yang telah memberikan petunjuk dan kemudahan untuk menyelesaikan studiku, kupersembahkan karya kecilku ini untuk:

Ayah dan Ibu tercinta yang selalu mendidik, mendoakan, berkorban, dan hal lain yang tak dapatku ungkapkan dengan kata-kata

Adik ku tersayang

Dosen pembimbing dan penguji yang sangat berjasa dan tidak lelah memberikan arahan serta masukan sehingga peulis dapat menyelesaikan tesisku

Sahabat dan teman-temanku, Terimakasih atas kebersamaan, do’a dan semangat yang selalu kalian berikan kepadaku.

Universitas Lampung

SANWACANA

Alhamdulillahi Robbil ‘alamin, Puji dan syukur Penulis ucapkan kepada Allah SWT, yang selalu melimpahkan rahmat dan kasih sayang-Nya, sehingga Penulis dapat menyelesaikan tesis ini. Sholawat serat salam senantiasa tetap tercurah kepada Nabi Muhammad SAW, tuntunan dan tauladan utama bagi seluruh umat manusia.

Tesis dengan judul “Algoritma PAM (*Partitioning Around Medoids*)-*Lite* dengan Program R untuk Data Berukuran Besar” adalah salah satu syarat untuk memperoleh gelar Magister Matematika di Universitas Lampung.

Dalam menyelesaikan tesis ini, banyak pihak yang telah membantu Penulis dalam memberikan bimbingan, dorongan, dan saran-saran. Sehingga dengan segala ketulusan dan kerendahan hati pada kesempatan ini Penulis mengucapkan terimakasih yang sebesar-besarnya kepada:

1. Dr. Khoirin Nisa, S.Si., M.Si., selaku Dosen Pembimbing 1 dan Dosen Pembimbing Akademik yang senantiasa memberikan bimbingan, saran, motivasi, nasehat serta masukan sehingga penulis dapat menyelesaikan perkuliahan dan tesis ini
2. Ir. Warsono, M.S., Ph.D., selaku Dosen Pembimbing 2 yang telah memberikan masukan dan saran dalam penyelesaian tesis
3. Prof. Drs. Mustofa Usman, M.A., Ph.D., selaku Dosen Pembahas 1 yang telah memberikan kritik dan saran kepada penulis dalam penyelesaian tesis
4. Prof. Dr. La Zakaria, S.Si., M.Sc., selaku Dosen Pembahas 2 yang telah memberikan kritik dan saran kepada penulis dalam penyelesaian tesis

5. Dr. Aang Nuryaman, S.Si., M.Si., selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam
6. Dr. Asmiati, S.Si., M.Si. selaku Ketua Program Studi Magister Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam
7. Dr. Eng. Suripto Dwi Yuwono, M.T. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung
8. Dosen, staf dan karyawan Jurusan Matematika FMIPA Universitas Lampung yang telah memberikan ilmu dan bantuan kepada penulis
9. Ayah dan Ibu yang tidak pernah lelah memberikan do'a, dukungan, kasih sayang, dan pengorbanan kepada Penulis
10. Adik ku tersayang
11. Teman-teman Magister Matematika Angkatan 2020.

Penulis juga menyadari bahwa dalam penulisan tesis ini masih banyak terdapat kekurangan. Oleh karena itu, Penulis mengharapkan saran dan kritik yang membangun guna penelitian selanjutnya agar lebih baik.

Bandar Lampung, 01 Juli 2022
Penulis,

Rizki Agung Wibowo

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR TABEL

Tabel Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Di era revolusi industri 4.0 kehadiran *big data* (data berukuran besar, baik data yang terstruktur maupun data yang tidak terstruktur) sering ditemukan di berbagai sektor industri dan sektor-sektor yang lain. Metode analisis statistik tradisional perlu ditingkatkan karena metode tersebut hanya efektif digunakan untuk data berukuran kecil, sehingga metode baru diperlukan agar kompatibel terhadap data berukuran besar dengan mempertahankan kemampuan untuk menghasilkan hasil analisis yang akurat. Salah satu analisis statistik tradisional yang perlu ditingkatkan yaitu analisis klaster, beberapa algoritma dan metode-metode analisis klaster telah dikembangkan dan sampai saat ini pun masih terus berkembang.

Analisis klaster merupakan teknik analisis multivariat yang memiliki tujuan untuk mengelompokkan objek-objek didalam suatu klaster dengan ragam seminimal mungkin sedangkan ragam antar klaster besar. Analisis klaster sering digunakan dalam studi *machine learning*, pengenalan pola (*pattern recognition*) dan statistik. Analisis klaster dibagi menjadi 4 sub-kategori yaitu partisi, hirarki, pengukuran jarak dan berbasis grid. Adapun metode partisi yang populer digunakan oleh para peneliti dan pengguna adalah k-medoid dan k-means.

K-medoid atau dapat disebut sebagai PAM (*partitioning around medoid*) adalah metode analisis klaster yang menjadi solusi efektif untuk menggantikan metode k-means ketika dihadapkan dengan data yang memiliki nilai pencilan, karena metode k-means dalam prosesnya menggunakan centroid (nilai tengah berupa rata-rata) sehingga akan menyebabkan hasil analisis yang kurang tepat. Sedangkan PAM dalam prosesnya menggunakan medoid sehingga dapat digunakan pada data yang

memiliki nilai pencilan atau dalam hal ini *robust* terhadap data pencilan. Medoid adalah objek yang letaknya terpusat di dalam suatu klaster.

Metode PAM memiliki keunggulan yaitu *robust* terhadap data pencilan, akan tetapi metode ini memiliki kelemahan dalam hal tingkat proses komputasi yang buruk pada data berukuran besar. Beberapa alternatif algoritma untuk mengatasi kelemahan pada PAM telah dikembangkan dan diperkenalkan, seperti CLARA (*The Clustering LARge Applications*) yang diperkenalkan oleh Kaufman dan Rousseeuw pada tahun 1990, kemudian CLARANS (*Clustering Large Application based on RANdomized Search*) yang diperkenalkan oleh Ng dan Han pada tahun 2002, kemudian Schubert dan Rousseeuw pada tahun 2018 memodifikasi PAM, CLARA dan CLARANS untuk meningkatkan waktu komputasi ketiga algoritma tersebut, dan pada tahun 2019 Peter et al. memperkenalkan analisis klaster PAM-*Lite* yang memiliki keunggulan berupa hasil analisis yang sangat mendekati seperti PAM, bahkan menghasilkan kelompok-kelompok yang lebih baik dibandingkan PAM. PAM-*Lite* juga memiliki keunggulan penghematan waktu yang signifikan seiring dengan meningkatnya ukuran dataset jika dibandingkan dengan CLARA, keunggulan diatas disimpulkan berdasarkan pengujian pemrograman komputer algoritma k-medoid PAM-*Lite* pada program MATLAB.

Pemrograman algoritma analisis statistika pada komputer sangat diperlukan untuk memudahkan pengguna dalam proses analisis. Program R merupakan salah satu bahasa pemrograman pada komputer yang sering digunakan oleh pengguna untuk melakukan proses analisis statistik, grafik dan simulasi data yang bersifat gratis dan *open source*. Terdapat banyak *script, function* dan *package* yang disediakan pada R yang dikembangkan oleh para pengembang yang dapat digunakan untuk proses analisis statistik. *Script, function* dan *package* analisis klaster metode PAM pun tersedia pada bahasa pemrograman ini.

Package yang disediakan pada program R yang digunakan untuk analisis klaster telah tersedia dan jumlahnya terus bertambah, sebagai contoh *stats* yang didalamnya mengandung *function* untuk analisis klaster hirarki (*single linkage,*

average linkage, *complete linkage* dan Ward), *cluster* yang didalamnya mengandung *function* untuk analisis klaster partisi (PAM, CLARA, CLARANS) dan lain-lain, sedangkan *package* yang didalamnya mengandung *function* untuk analisis klaster PAM-*Lite* belum tersedia pada program R. *Function* analisis klaster PAM-*Lite* belum tersedia di program R dikarenakan belum ada pengembang yang membuat *function* tersebut, hal ini sangat disayangkan mengingat PAM-*Lite* diklaim memiliki keunggulan berupa hasil analisis yang sangat mendekati seperti PAM, bahkan menghasilkan kelompok-kelompok yang lebih baik dibandingkan PAM, juga memiliki keunggulan penghematan waktu yang signifikan seiring dengan meningkatnya ukuran dataset jika dibandingkan dengan CLARA, oleh karena itu penulis tertarik untuk membuat *function* PAM-*Lite* pada program R dan menguji keefektifannya menggunakan data simulasi sebagai topik penelitian tesis ini.

1.2 Rumusan Masalah

Penelitian yang dilakukan pada tesis ini adalah penelitian yang akan menjawab pertanyaan:

1. Bagaimana cara mengkonstruksi algoritma PAM-*Lite* pada program R ?
2. Bagaimana menyelesaikan analisis klaster pada data yang besar menggunakan algoritma PAM-*Lite* pada program R ?
3. Bagaimana hasil evaluasi analisis klaster PAM-*Lite* pada data yang besar ?.

1.3 Batasan Masalah

Data yang digunakan pada penelitan ini adalah data dua dimensi bertipe data diskrit dengan banyaknya objek antara 1024 hingga 100000.

1.4 Tujuan Penelitian

Penelitian ini bertujuan mengkonstruksi sebuah *function* PAM-*Lite* pada program R dan menguji keefektifannya menggunakan data simulasi.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah:

1. Penelitian ini akan menambah kejelasan kepada pembaca bagaimana menyelesaikan analisis klaster pada data yang besar menggunakan algoritma PAM-*Lite* pada program R
2. Memberikan sumbangan pemikiran dalam rangka memperluas dan memperdalam pengetahuan ilmu statistika khususnya mengenai analisis klaster algoritma PAM-*Lite* pada program R
3. Dapat dijadikan referensi untuk pengaplikasikan analisis klaster PAM-*Lite* pada program R.

II. TINJAUAN PUSTAKA

Pada bab ini akan dibahas definisi tentang analisis klaster PAM-*Lite*, R-*Square*, lebar *silhouette* dan program R.

2.1 Analisis Klaster

Analisis klaster merupakan teknik analisis multivariat yang memiliki tujuan untuk mengelompokkan objek-objek didalam suatu klaster dengan ragam seminimal mungkin sedangkan ragam antar klaster sebesar mungkin (Härdle and Simar, 2007). Analisis klaster didasarkan pada jarak untuk merepresentasikan ukuran kesamaan dan ukuran yang sering digunakan adalah jarak Euclid (Härdle and Simar, 2007; Margaritis *et al*., 2020).

2.1.1 Nilai Z (*Z-Score*)

Nilai Z adalah selisih antara sebuah nilai X dan rata-rata (μ) dibagi dengan standar deviasinya (σ) (Lind *et al.*, 2021; McClave *et al.*, 2022). Nilai Z dapat didefinisikan dalam bentuk rumus:

$$Z = \frac{X - \mu}{\sigma} \tag{2.1}$$

keterangan:

X : nilai beberapa pengamatan atau pengukuran tertentu

μ : rata-rata distribusi

σ : standar deviasi distribusi

2.1.2 Ukuran Kemiripan (Jarak)

Analisis klaster adalah proses mengelompokkan data yang yang memiliki kesamaan atau kedekatan antar satu data dengan data yang lain yang membentuk klaster atau kelompok sehingga kelompok tersebut memiliki tingkat kesamaan yang tinggi sedangkan antar klaster memiliki tingkat kesamaan yang kecil (Faisal *et al.*, 2020; Nishom, 2019).

Pada analisis klaster digunakan jarak Euclid sebagai alat ukur kemiripan, yang didefinisikan sebagai berikut:

$$d(x,y) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})'(\boldsymbol{x}-\boldsymbol{y})} = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2} \tag{2.4}$$

Keterangan:

$\boldsymbol{x}$: vektor objek x

$\boldsymbol{y}$: vektor objek y

p : banyaknya variabel

(Härdle and Simar, 2007)

Atau dapat ditulis sebagai berikut:

$$d(i,j) = d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{2.5}$$

Keterangan :

d_{ij} : Jarak antara objek ke-i dan obyek ke-j

x_{ik} : data dari objek ke-i pada variabel ke-k

x_{jk} : data dari objek ke-j pada variabel ke-k

Jarak Euclid digunakan dalam analisis klaster PAM dikarenakan jarak Euclid memiliki tingkat efektifitas yang tinggi dalam perhitungan data dengan dimensi yang kecil (Mohibullah *et al.*, 2015), pada analisis klaster metode K-Means jarak Euclid juga unggul dalam hal banyaknya iterasi yang dilakukan selama perhitungan centroid dibandingkan jarak lainnya seperti jarak Manhattan (Sinwar and Kaushik, 2014) dan jarak Minkowski (Nishom, 2019)

2.1.3 Analisis Klaster K-Medoid

K-Medoid atau *Partitioning Around Medoids* (PAM) diperkenalkan oleh Kaufman dan Rousseuw, merupakan metode analisis klaster yang mirip dengan K-Means. PAM adalah teknik pengelompokan data secara partisi klasik dengan menentukan banyaknya klaster yang akan dibentuk, konsep dasar dari analisis klaster PAM adalah untuk membentuk sebanyak k klaster pada n objek yang langkah awalnya adalah memilih secara acak objek representatif (medoid) untuk setiap klaster (Gultom *et al.*, 2018; Kaur *et al.*, 2014).

Metode PAM memiliki kelebihan untuk mengatasi kelemahan pada K-Means yang sensitif terhadap nilai pencilan, dimana pencilan adalah objek dengan nilai yang besar sehingga memungkinkan menyimpang dari distribusi data (Larasati *et al.*, 2021; Olukanmi *et al.*, 2019; Park and Jun, 2009). Adapun algoritma analisis klaster PAM (Gultom *et al.*, 2018; Park and Jun, 2009) sebagai berikut:

Langkah 1: (Memilih inisial medoid)

1-1 Hitung jarak diantara setiap pasangan semua objek berdasarkan ukuran jarak yang dipilih (dalam hal ini jarak Euclid)

1-2 Hitung v_j untuk objek j dengan:

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \; ; \qquad j = 1,2,3, \dots, n$$

1-3 Urutkan v_j secara meningkat (dari terkecil hingga terbesar). Pilih k objek pertama v_j terkecil sebagai inisial medoid

1-4 Dapatkan hasil klaster awal dengan menetapkan setiap poin objek ke medoid terdekat

1-5 Hitung jumlah jarak dari semua objek terhadap medoidnya

Langkah 2: Perbarui medoid

Temukan medoid baru untuk setiap klaster, yaitu objek yang meminimalkan jumlah jarak ke objek lain dalam klasternya. Perbarui medoid lama disetiap klaster dengan medoid baru

Langkah 3: Mengelompokan objek terhadap medoid

3-1 Mengelompokan setiap objek ke medoid terdekat

3-2 Hitung jumlah jarak dari semua objek terhadap medoidnya. Jika jumlah jarak sama dengan langkah sebelumnya, maka hentikan algoritma. Sebaliknya, maka lakukan langkah 2.

2.1.4 Analisis Klaster PAM-*Lite*

Pada tahun 2019, algoritma PAM-*Lite* diperkenalkan oleh Peter O. Olukanmi, Fulufhelo Nelwamondo dan Tshilidzi Marwala, algoritma PAM-*Lite* merupakan pengembangan dari PAM klasik. Algoritma ini memiliki kelebihan berupa efisiensi dibandingkan PAM klasik, bahkan dapat menghasilkan kualitas pengelompokan yang lebih baik daripada PAM (Olukanmi *et al.*, 2019). Algoritma PAM-*Lite* didasarkan pada paradigma K-Means-*Lite* yang akan dijelaskan pada sub-bab berikutnya

2.1.4.1 Paradigma K-Means-*Lite*

K-Means-*Lite* diperkenalkan oleh Peter O. Olukanmi, Fulufhelo Nelwamondo dan Tshilidzi Marwala pada tahun 2018. Cara kerja K-Means-*Lite* adalah menggabungkan teorema limit pusat (CLT) dengan metode K-Means. CLT memberikan dasar untuk menarik kesimpulan tentang populasi menggunakan sampel acak. Jika diberikan populasi P yang memiliki rata-rata μ dan standar deviasi σ, maka distribusi dari rata-rata sampel $\bar{X}_i$ $(i = 1, \dots, n)$ dimana n adalah sampel berukuran s yang dipilih secara acak dari populasi P yang cenderung akan berdistribusi normal yang memiliki mean μ dan standar deviasi $\sigma/\sqrt{s}$ dengan $n, s \to \infty$. Pertama, jika populasi yang berukuran besar yang akan dipartisi, dan banyaknya klaster $k = 1$, maka μ bertindak sebagai centroid. Dengan kata lain CLT menunjukan bahwa μ adalah centroid dari populasi, dapat diperkirakan dengan terlebih dahulu mengelompokan masing-masing dari n sampel dengan ukuran yang sama s kedalam $k = 1$ klaster, yang menghasilkan rata-rata $\bar{X}_i$ sebagai centroid disetiap sampel dan menggabungkan n rata-rata tersebut menjadi

satu dataset, yang ketika dikelompokan menjadi $k = 1$ klaster akan menghasilkan centroid c yang mendekati μ, apabila s dan n membesar. Dengan demikian pendekatan berbasis inferensia ini adalah cara alternatif yang lebih efisien untuk menghitung centroid. Teorema 1 akan dijelaskan untuk banyaknya klaster k secara umum (Olukanmi *et al.*, 2018, 2019).

Teorema 1

Diberikan dataset P yang terdiri dari k klaster yang memiliki centroid $\mu_1, \dots, \mu_k$. Jika sampel acak n berukuran s diambil dari P, dan masing-masing sampel dipartisi menjadi k klaster yang didefinisikan sebagai k centroid $\bar{X}_{i1}, \dots, \bar{X}_{ik}$; $i = 1, \dots, n$ maka k centroid diperoleh dengan mempartisi dataset $\bar{X}_{ij}, (i = 1, \dots, n\,; j = 1, \dots, k)$ menjadi k klaster, mendekati $\mu_1, \dots, \mu_k$, dengan $n \to \infty$ (Olukanmi *et al.*, 2018, 2019).

Bukti Teorema 1

Pertimbangkan k dataset $D_1, \dots, D_k$, memiliki distribusi dan rata-rata $\mu_1, \dots, \mu_k$ yang berbeda. Jika digabungkan menjadi satu dataset P, maka sampel berukuran $s >> k$ yang kemungkinan akan berisi poin dari masing-masing $D_1, \dots, D_k$, dan kemungkinannya akan meningkat jika $s \to \infty$. Ketika sampel dipilih, titik-titik dari setiap distribusi dapat dikelompokan secara bersamaan melalui proses pengklasteran, menghasilkan centroid (rata-rata sampel) untuk setiap klaster. Jika prosedur pengklasteran diulang sebanyak n kali, maka akan didapatkan kumpulan nk rata-rata sampel. Jika kumpulan rata-rata sampel tersebut dipartisi menjadi k klaster, berdasarkan CLT, rata-rata atau centroid dari kelompok ini mendekati $\mu_1, \dots, \mu_k$, yang mendefinisikan solusi K-Means untuk kombinasi dataset P (Olukanmi *et al.*, 2018, 2019) ■.

2.1.4.2 Algoritma PAM-*Lite*

Pendekatan inferensia pada sub-bab 2.1.4.1 untuk K-Means juga akan diaplikasikan pada PAM. Dasar dari kedua metode ini adalah centroid dan medoid merupakan estimasi dari pusat klaster. Oleh karena itu, algoritma PAM-

Lite menggunakan pendekatan inferensia yang identik dengan K-Means-*Lite*. Perbedaan antara kedua algoritma tersebut adalah pada algoritma PAM-*Lite*, sampel dan kumpulan pusat klaster gabungan yang diperoleh dari sampel dikelompokkan menggunakan algoritma PAM tidak seperti K-Means-*Lite* yang didasarkan pada K-Means. Algoritma PAM-*Lite* dapat dijelaskan sebagai berikut:

1. Pilih n sampel untuk setiap ukuran s dari dataset P
2. Jalankan algoritma PAM pada setiap sampel dan simpan medoid di P'
3. Jalankan algoritma PAM pada P' untuk mendapatkan set medoid C
4. Tetapkan setiap titik di P ke medoid terdekat di C.

Untuk memilih ukuran sampel s, secara empirik mengikuti ukuran pada K-Means-*Lite* yaitu mengikuti ukuran sampel pada algoritma CLARA (Kaufman and Rousseeuw, 1990), yaitu 5 sampel dengan setiap sampel berukuran $40 + 2k$. Adapun *pseudocode* algoritma PAM-*Lite* (Olukanmi *et al.*, 2019) sebagai berikut:

Input :

P adalah dataset yang memiliki N poin

k adalah banyaknya klaster yang diinginkan

n adalah banyaknya sampel yang dipilih

s ukuran yang dipilih untuk setiap sampel (dalam hal ini $40 + 2k$)

Output :

A adalah N-vektor alamat klaster

C adalah himpunan k medoid klaster

Prosedur :

D = NULL

untuk i = 1 hingga n

S_i = sampel acak (P, s)

(a_i, c_i) = PAM(S_i, k)

D = gabungkan secara baris(D, c_i)

selesai

C = PAM(D, k)

A = Tetapkan setiap poin terhadap medoid terdekat(P, C)

2.2 Nilai *R-Square*

Untuk mengukur kualitas dari suatu model statistik sering diukur dengan menggunakan proporsi yang dapat dijelaskan oleh model, adapun statistik yang digunakan untuk mengukur hal tersebut adalah R^2. Pada analisis regresi berganda, statistik R^2 (koefisien determinasi) didefinisikan sebagai besaran variasi dari variabel terikat yang dapat dijelaskan oleh model (Akossou and Palm, 2013).

Pada analisis klaster, R^2 didefinisikan sebagai ukuran yang merepresentasikan seberapa dekat objek-objek pada klaster yang sama. Klaster yang optimal dikaitkan dengan memaksimalkan nilai R^2 yaitu dengan memaksimalkan varian antar klaster dan meminimalkan varian dalam klaster (Loperfido and Tarpey, 2018).

Indeks R-*square* dapat didefinisikan sebagai berikut:

$$RS = \frac{SS_B}{SS_T} = \frac{SS_T - SS_W}{SS_T}$$

$$RS = \frac{\left\{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2\right\} - \left\{\sum_{k=1}^{c}\sum_{j=1}^{p}\sum_{i=1}^{n_c}\left(x_{ijk} - \bar{x}_{jk}\right)^2\right\}}{\left\{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2\right\}} \tag{2.6}$$

Keterangan:

x_{ij} : objek ke – i pada variabel j

$\bar{x}_j$: rata-rata pada variabel j

$\bar{x}_{jk}$: rata-rata variabel j pada klaster k

x_{ijk} : objek ke – i pada klaster k pada variabel j

p : banyaknya variabel

c : banyaknya klaster

n_c : banyaknya objek pada klaster c

Semakin besar nilai R-*square* maka klaster yang dihasilkan akan semakin baik. Nilai R-*square* memiliki rentang nilai dari nol sampai satu (Loperfido and Tarpey, 2018; Sharma, 1996).

2.3 Lebar *Silhouette*

Lebar *silhouette* adalah rata-rata dari nilai *silhouette* setiap observasi. Nilai *silhouette* mengukur derajat kepercayaan didalam sebuah klaster, dengan kategori klaster yang baik jika nilainya mendekati 1 dan buruk jika nilainya mendekati -1.

Untuk pengamatan i nilai *silhouette* didefinisikan sebagai berikut:

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{2.7}$$

$$a_i = \frac{1}{n\big(C(i)\big) - 1} \sum_{j \in C(i)} dist(i,j)$$

$$b_i = \min_{C_k \in \mathcal{C} \setminus C(i)} \sum_{j \in C_k} \frac{dist(i,j)}{n(C_k)}$$

dengan a_i adalah rata-rata jarak antara i dan semua observasi pada klaster yang sama, b_i adalah rata-rata jarak antara i dan observasi pada "klaster tetangga terdekat", $C(i)$ adalah klaster yang berisi observasi ke-i, $dist(i,j)$ adalah jarak antar observasi i dan j (jarak Euclid, Manhattan), $n(C)$ adalah kardinalitas (jumlah elemen dalam satu set klaster) dari klaster C (Brock *et al.*, 2008).

2.4 Program R

R adalah bahasa pemrograman yang dikhususkan untuk komputasi statistik dan grafik yang bersifat gratis dan *open source*. Nama R diambil berdasarkan inisial kedua pemiliknya yaitu Robert Gentleman dan Ross Ihaka (Khan, 2013). Program R rilis perdana pada tahun 1993. R tersedia pada semua sistem operasi komputer sehingga R lebih fleksibel penggunaannya jika dibandingkan dengan program statistika berbayar lainnya (Ozgur *et al.*, 2021). R adalah program yang sangat fleksibel dengan banyak fitur tambahan yang dapat diunduh dalam bentuk *package*.

R adalah rangkaian perangkat lunak terintegrasi untuk melakukan manipulasi data, kalkulasi dan visualisasi grafik (Khan, 2013). R memiliki banyak keunggulan salah satu diantaranya adalah waktu proses perhitungan yang cepat, bersifat gratis, ahli statistika dapat mengembangkan metode analisis statistik dengan membuat *package*, kemampuan dalam visualisasi grafik yang baik, unggul dalam melakukan simulasi yang memerlukan perhitungan secara intensif (Nisa dan Wibowo, 2021).

2.4.1 *Function* pada R

Perintah untuk membuat fungsi pada R adalah sebagai berikut:

```
Nama_function = function(argumen_1, argumen_2, ...)
{
 Statement #operasi yang diinginkan
 return(objek) #berfungsi untuk mengeluarkan objek
}
```

(Dinov, 2018; Nisa dan Wibowo, 2021).

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan pada semester genap tahun ajaran 2021/2022 bertempat di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

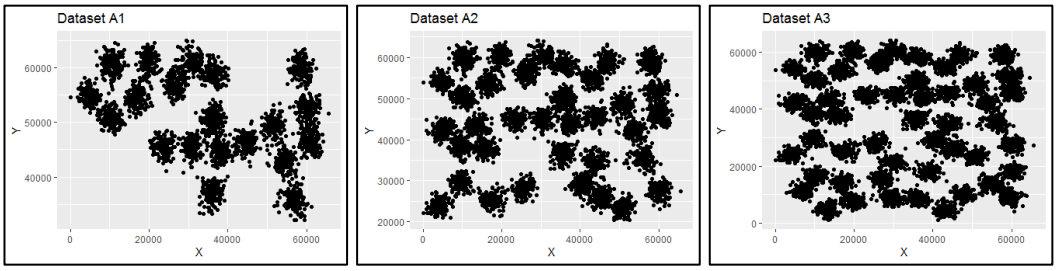
3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data dua dimensi bertipe data diskrit yang digunakan pada paper karya Pasi Fränti dan Sami Sieranoja tahun 2018, yaitu berupa enam data *basic clustering benchmark* (Fränti and Sieranoja, 2018) sebagai berikut:

Tabel 1. Informasi *basic clustering benchmark*

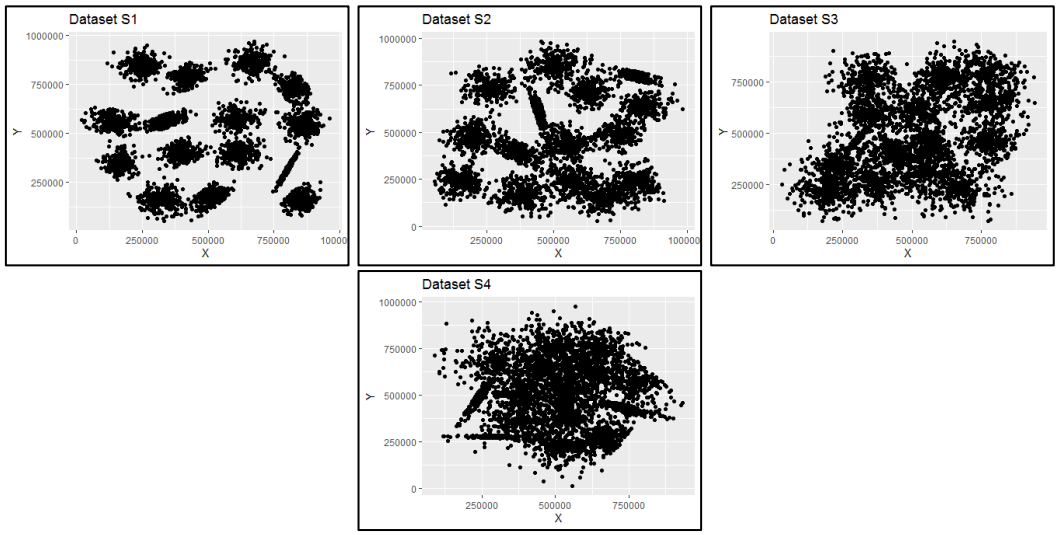
Sumber	**Data**	**Ukuran** $(\boldsymbol{N})$	**Banyak Klaster** $(\boldsymbol{k})$	**Banyak Objek Per klaster**
Kärkkäinen and Fränti (2002)	A	3000-7500	20,30,50	150
Fränti and Virmajoki (2006)	S	5000	15	333
Fränti *et al.* (2006)	Dim	1024	16	64
Fränti *et al.* (2016)	G2	2048	2	1024
Zhang *et al.* (1997)	Birch	100000	100	1000
Rezaei and Fränti (2016)	Unbalance	6500	8	100-2000

- **A sets**: data ini berisi set klaster dua dimensi berbentuk bulat dengan banyaknya klaster masing-masing A1, A2, A3 yaitu $k = 20, 35, 50$



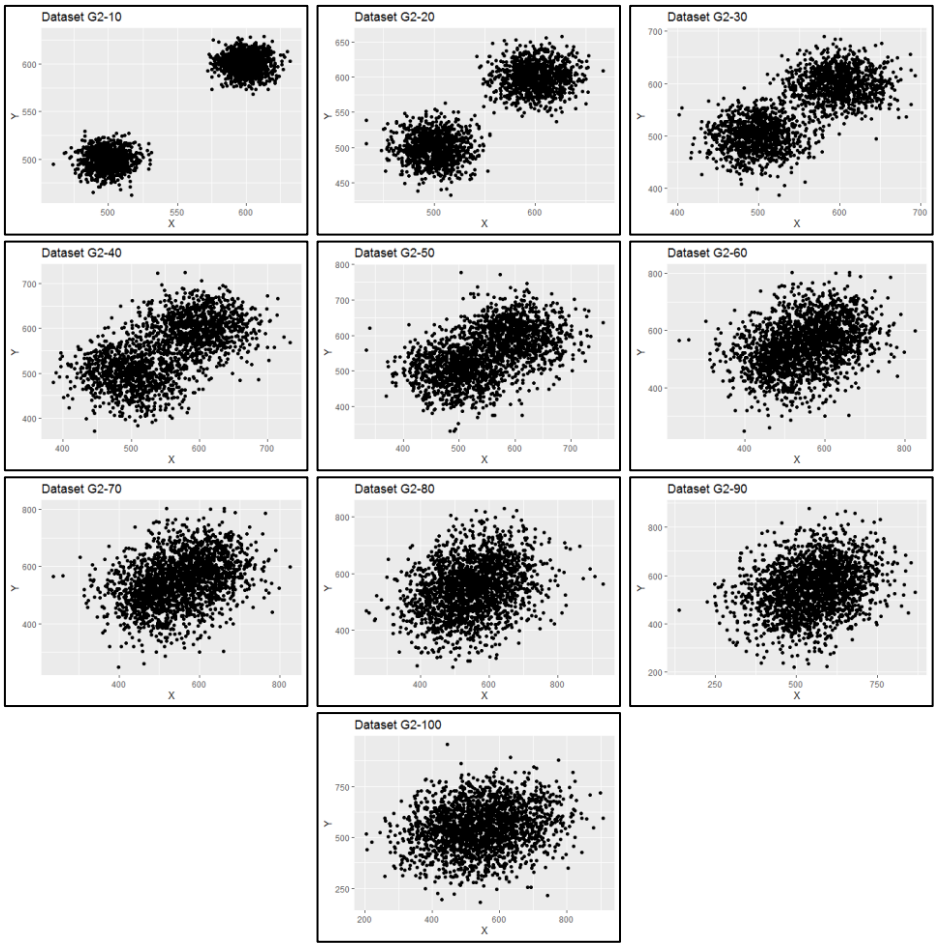
Gambar 1. Plot A Sets.

- **S sets**: data ini berisi klaster Gaussian dengan *overlap* (pemisahan klaster) klaster yang bervariasi mulai dari 9% hingga 44%, banyaknya klaster $k = 15$ dan $N = 5000$. Kebanyakan klaster berbentuk bulat, tetapi beberapa klaster tidak berbentuk bulat. Set terakhir yaitu S4 memiliki *overlap* yang tinggi



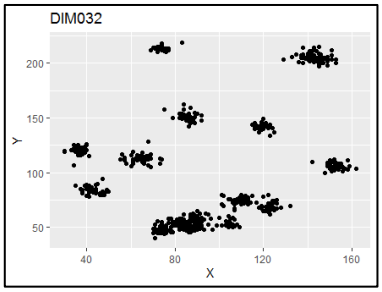
Gambar 2. Plot S Sets.

- **G2**: data ini berisi dua klaster Gaussian pada lokasi titik tetap, setiap klaster berisi 1024 poin. *Overlap* diatur dengan cara memperbesar standar deviasi distribusi Gaussian dari 10 hingga 100.



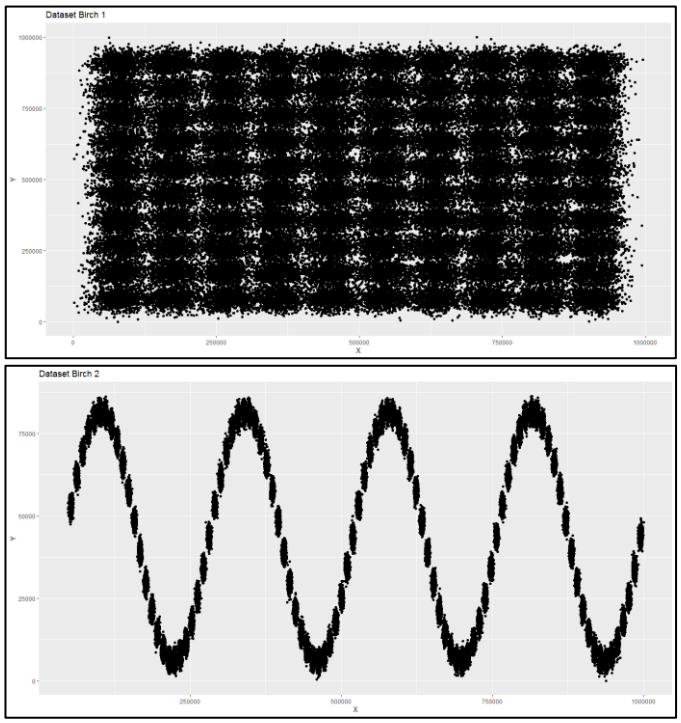
Gambar 3. Plot G2.

- **DIM**: data ini berisikan klaster yang terpisah dengan baik pada ruang dimensi tinggi dengan dimensi yang bervariasi mulai dari 32 hingga 1024. Titik setiap klasternya bersifat acak dan diambil dari distribusi Gaussian. Pada penelitian ini hanya digunakan DIM032 dengan 2 variabel pertama yaitu data dengan banyaknya klaster $k = 16$ dan $N = 1024$, Gambar 4 adalah plot dua dimensi dua kolom pertama dari data DIM032



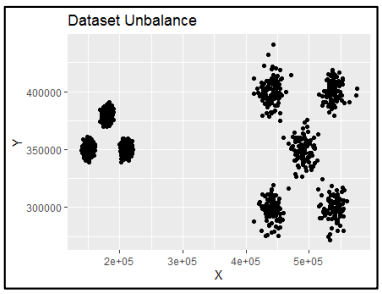
Gambar 4. Plot dua dimensi DIM032.

- **Birch**: data ini diperkenalkan oleh Tian Zhang, Raghu Ramakrishnan dan Miron Livny pada tahun 1997 (Zhang *et al.*, 1997). Pada penelitian ini digunakan 2 data besar yaitu Birch 1 dan Birch 2 yang memiliki bentuk klaster bulat dengan $k = 100$ dan $N = 100.000$. Birch 1 memiliki bentuk berupa kisi-kisi (*grid*) 10×10 sedangkan Birch 2 memiliki bentuk fungsi kurva sin



Gambar 5. Plot Birch 1 (atas) dan Birch 2 (bawah).

- **Unbalance**: data ini memiliki delapan klaster dalam dua grup yang terpisah dengan baik. Tiga klaster pertama terdiri atas 2000 poin per klaster dan lima lainnya terdiri atas 100 poin tiap klasternya.

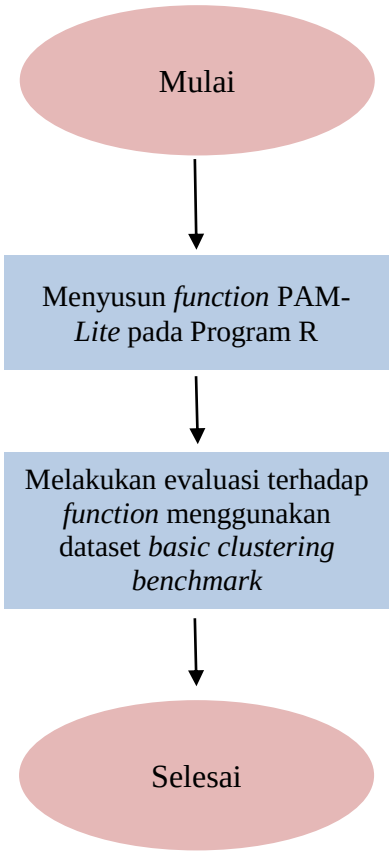


Gambar 6. Plot Unbalance.

3.3 Metode Penelitian

Penelitian ini dilakukan secara studi pustaka yaitu mempelajari buku-buku teks, jurnal serta akses internet yang menunjang proses penelitian. Penelitian ini menggunakan program Rstudio versi 1.4.1103 yang dijalankan pada laptop dengan spesifikasi *processor* Intel Core i3-4030U @ 1,90 GHz, ram 4 GB dan sistem operasi Windows 8.1 Pro. Adapun langkah-langkah penelitian yang dilakukan sebagai berikut:

1. Menyusun *function* PAM-*Lite* pada program R sesuai dengan algoritma
2. Melakukan evaluasi terhadap *function* yang telah disusun pada langkah 1 menggunakan enam data *basic clustering benchmark*, statistik yang digunakan untuk evaluasi adalah R-*square* dan lebar *silhouette*.



Gambar 7. Alur Penelitian

V. KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil dan pembahasan yang telah dijelaskan pada Bab IV, maka dapat diambil kesimpulan sebagai berikut:

1. *Function* PAM-*Lite* yang telah dikonstruksi dapat diproses dengan baik pada program R
2. Berdasarkan data *basic clustering benchmark*, *function* `pam.lite` efektif digunakan pada data berukuran besar hingga 7500 sedangkan untuk data berukuran 100000 dapat menggunakan *function* `pam.lite.2` yang merupakan modifikasi dari `pam.lite`
3. Dengan hanya menggunakan 5 sampel acak yang setiap sampelnya berukuran $40 + 2k$, berdasarkan R-*Square*, lebar *silhouette* dan waktu proses, dapat disimpulkan bahwa PAM-*Lite* lebih efisien jika dibandingkan dengan PAM pada data *basic clustering benchmark*.

5.2 Saran

Untuk penelitian lanjutan disarankan untuk mengkaji algoritma PAM-*Lite* menggunakan statistik evaluasi lain seperti Dunn *index*, *Connectivity* dan Rand *index*, Calinski-Harabasz *index*, dan Hartigan *index* serta mengkajinya menggunakan data bangkitan berukuran lebih dari 7500 dengan variabel saling berkorelasi kemudian dipadukan dengan analisis komponen utama untuk mengatasi data tersebut apabila ukuran jarak yang dipakai pada algoritma PAM-*Lite* adalah jarak Euclid.

DAFTAR PUSTAKA

Akossou, A. Y. J., and Palm, R. 2013. Impact of Data Structure on the Estimators R-Square and Adjusted R-Square in Linear Regression. *International Journal of Mathematics & Computation*, *20*(3).

Brock, G., Pihur, V., Datta, S., and Datta, S. 2008. clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, *25*(4 SE-Articles), 1–22.

Dinov, I. D. 2018. *Data science and predictive analytics: Biomedical and health applications using R*. Springer International Publishing.

Faisal, M., Zamzami, E. M., and Sutarman. 2020. Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance. *Journal of Physics: Conference Series*, *1566*(1).

Fränti, P., Mariescu-Istodor, R., and Zhong, C. 2016. XNN Graph. *IAPR Joint Int. Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, *LNCS 10029*, 207–217.

Fränti, P., and Sieranoja, S. 2018. K-means properties on six clustering benchmark datasets. *Applied Intelligence 2018*, *48*(12), 4743–4759.

Fränti, P., and Virmajoki, O. 2006. Iterative shrinking method for clustering problems. *Pattern Recognition*, *39*(5), 761–765.

Fränti, P., Virmajoki, O., and Hautamäki, V. 2006. Fast Agglomerative Clustering Using a K-Nearest Neighbor Graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(11), 1875–1881.

Gultom, S., Sriadhi, S., Martiano, M., and 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).

Härdle, W. K., and Simar, L. 2007. Applied Multivariate Statistical Analysis. In *Applied Multivariate Statistical Analysis* (2 ed.). Springer-Verlag.

Kaufman, L., and Rousseeuw, P. J. 1990. Clustering Large Applications (Program CLARA). In *Wiley Series in Probability and Statistics* (pp. 126–163). John Wiley & Sons, Ltd.

Kaur, N. K., Kaur, U., and Singh, D. 2014. K-Medoid clustering algorithm- A review. *International Journal of Computer Application and Technology*, *1*(1), 42–45.

Kärkkäinen, I., and Fränti, P. 2002. *Dynamic local search algorithm for the clustering problem*.

Khan, A. M. 2013. R-software: A Newer Tool in Epidemiological Data Analysis. *Indian Journal of Community Medicine : Official Publication of Indian Association of Preventive & Social Medicine*, *38*(1), 56.

Larasati, S. D. A., Nisa, K., and Herawati, N. 2021. Robust Principal Component Trimmed Clustering of Indonesian Provinces Based on Human Development Index Indicators. *Journal of Physics: Conference Series*, *1751*(1), 0–8.

Lind, D., Marchal, W., and Wathen, S. 2021. *Statistical Techniques in Business and Economics* (18 ed.). Mc Graw Hill.

Loperfido, N., and Tarpey, T. 2018. Some remarks on the R2 for clustering. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *11*(3), 135–148. https://doi.org/10.1002/SAM.11378

Margaritis, A., Soenen, H., Fransen, E., Pipintakos, G., Jacobs, G., Blom, J., and Van den bergh, W. 2020. Identification of ageing state clusters of reclaimed asphalt binders using principal component analysis (PCA) and hierarchical cluster analysis (HCA) based on chemo-rheological parameters. *Construction and Building Materials*, *244*, 1–21.

McClave, J. T., Benson, P. G., and Sincich, T. T. 2022. *Statistics for Business and Economics* (14 ed.). Pearson Education Limited.

Mohibullah, M., Hossain, M. Z., and Hasan, M. 2015. Comparison of Euclidean Distance Function and Manhattan Distance Function Using K-Mediods. *International Journal of Computer Science and Information Security (IJCSIS)*, *13*(10), 61–71.

Nisa, K., dan Wibowo, R. A. 2021. *Simulasi Data Statistik Menggunakan R* (1 ed.). Teknosain.

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.

Olukanmi, P. O., Nelwamondo, F., and Marwala, T. 2018. K-Means-Lite: Real Time Clustering for Large Datasets. *5th International Conference on Soft Computing and Machine Intelligence, ISCMI 2018*, 54–59.

Olukanmi, P. O., Nelwamondo, F., and Marwala, T. 2019. PAM-lite: Fast and accurate k-medoids clustering for massive datasets. *Proceedings - 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa, SAUPEC/RobMech/PRASA 2019*, *typically 5*, 200–204. h

Ozgur, C., Colliau, T., Rogers, G., and Hughes, Z. 2021. MatLab vs. Python vs. R. *Journal of Data Science*, *15*(3), 355–372.

Park, H. S., and Jun, C. H. 2009. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341.

Rezaei, M., and Fränti, P. 2016. Set Matching Measures for External Cluster Validity. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 2173–2186.

Sharma, S. 1996. *Applied Multivariate Techniques*. John Wiley and Sons, Inc.

Sinwar, D., and Kaushik, R. 2014. Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, *2*(5), 270–274.

Zhang, T., Ramakrishnan, R., and Livny, M. 1997. BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery*, *1*(2), 141–182.