

**ANALISIS KLASTER PADA DATASET BESAR
DENGAN ALGORITMA CLARANS
(Studi Kasus : Tingkat Kemiskinan di 221 Kota/Kabupaten
di Indonesia Tahun 2020)**

SKRIPSI

Diajukan guna Memenuhi Sebagian dari Syarat Memperoleh Gelar
Sarjana Matematika (S.Mat) pada Program Studi Matematika



Oleh:

Fitri Nurkholifah

1700014

**PROGRAM STUDI MATEMATIKA
DEPARTEMEN PENDIDIKAN MATEMATIKA
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
2021**

Fitri Nurkholifah, 2021

*ANALISIS KLASTER PADA DATASET BESAR DENGAN ALGORITMA CLARANS (STUDI KASUS : TINGKAT KEMISKINAN
DI 221 KOTA/KABUPATEN DI INDONESIA TAHUN 2020)*

Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu

**ANALISIS KLASSTER PADA DATASET BESAR
DENGAN ALGORITMA CLARANS
(Studi Kasus : Tingkat Kemiskinan di 221 Kota/Kabupaten
di Indonesia Tahun 2020)**

LEMBAR HAK CIPTA

Oleh
Fitri Nurkholifah
1700014

Tugas akhir yang diajukan guna memenuhi sebagian syarat untuk memperoleh gelar Sarjana Matematika pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

© Fitri Nurkholifah 2021
Universitas Pendidikan Indonesia
Juli 2021

Hak Cipta dilindungi oleh undang – undang
Skripsi ini tidak boleh diperbanyak seluruh ataupun sebagian dengan dicetak ulang,
difotokopi, atau cara lainnya tanpa izin dari penulis

LEMBAR PENGESAHAN

FITRI NURKHOLIFAH

**ANALISIS KLASSTER PADA DATASET BESAR
DENGAN ALGORITMA CLARANS
(Studi Kasus : Tingkat Kemiskinan di 221 Kota/Kabupaten
di Indonesia Tahun 2020)**

Disetujui dan disahkan oleh pembimbing:

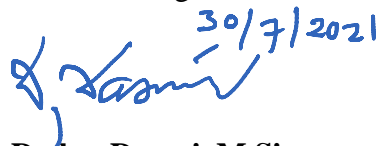
Pembimbing I



Hj. Dewi Rachmatin, S.Si., M.Si.

NIP. 196909291994122001

Pembimbing II



Dr. Dadan Dasari, M.Si.

NIP.196407171991021001

Mengetahui,
Ketua Departemen Pendidikan Matematika



Dr. H. Dadang Juandi, M.Si.

NIP. 196401171992021001

LEMBAR PERNYATAAN

Dengan ini saya menyatakan bahwa skripsi yang berjudul “Analisis Klaster pada Dataset Besar dengan Algoritma CLARANS (Studi kasus: Tingkat Kemiskinan di 221 Kota/Kabupaten di Indonesia Tahun 2020)” murni merupakan karya saya sendiri. Saya tidak melakukan tindakan yang melanggar hukum seperti melakukan penjiplakan atau pengutipan dengan cara – cara yang tidak sesuai dengan aturan yang berlaku dalam masyarakat keilmuan. Jika di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau adanya klaim dari pihak lain terhadap keaslian dari skripsi ini, maka saya siap untuk menanggung sanksi sesuai dengan peraturan yang berlaku.

Bandung, Juli 2021

Yang Membuat Pernyataan,



Fitri Nurkholifah

1700014

KATA PENGANTAR

Puji syukur mari panjatkan atas kehadiran Allah SWT karena berkat rahmat serta hidayah-Nya penulis masih diberikan kesehatan sehingga dapat menyelesaikan skripsi yang berjudul “Analisis Klaster pada Dataset Besar dengan Algoritma CLARANS (Studi kasus: Tingkat Kemiskinan di 221 Kota/Kabupaten di Indonesia Tahun 2020)” tepat pada waktunya. Skripsi ini merupakan salah satu syarat guna meraih gelar Sarjana Matematika (S.Mat) di Universitas Pendidikan Indonesia.

Terselesaikannya skripsi ini tidak lepas dari bimbingan serta dorongan dari berbagai pihak. Oleh karena itu, pada kesempatan kali ini penulis ingin mengucapkan terimakasih kepada pihak – pihak yang telah membantu dalam memberikan saran, kritik, serta motivasi sehingga skripsi ini dapat selesai dengan baik dan tepat waktu. Pihak – pihak tersebut ialah:

1. Ibu Hj. Dewi Rachmatin, S.Si., M.Si. selaku dosen pembimbing I yang selalu memberikan saran serta motivasi selama proses pengerjaan skripsi ini. Selain itu juga selalu mengoreksi setiap kekeliruan penulisan yang dilakukan penulis.
2. Bapak Dr. Dadan Dasari, M.Si. selaku dosen pembimbing II yang telah memberikan bimbingan serta saran selama proses penyusunan skripsi ini dimulai saat mata kuliah Kapita Selekta.
3. Ibu Dr. Khusnul Novianingsih, S.Si, M,Si. selaku dosen pembimbing akademik yang telah memberikan bimbingan serta arahan selama menempuh studi di Universitas Pendidikan Indonesia.
4. Semua dosen Program Studi Matematika Universitas Pendidikan Indonesia yang telah memberikan ilmu pengetahuan yang bermanfaat selama masa studi.
5. Orang tua, adik, serta keluarga besar yang selalu memberikan dorongan semangat serta doa agar penulis diberikan kelancaran selama proses pengerjaan skripsi ini.

6. Regina, Amira, Ahmad, Aulia, serta Indah selaku teman terdekat yang dengan sabar selalu mendengarkan keluh kesah penulis selama mengerjakan skripsi ini serta memberikan saran dan motivasi.
7. Pihak – pihak lainnya yang telah membantu dalam proses penyelesaian skripsi ini yang tidak dapat disebutkan satu persatu namanya oleh penulis.

Hanya ucapan terimakasih serta doa yang dapat penulis sampaikan. Semoga Allah SWT membalas semua kebaikan Ibu/Bapak. Penulis menyadari bahwa skripsi ini jauh dari kata sempurna. Oleh karena itu, saran serta kritik yang membangun sangat diperlukan agar skripsi ini menjadi lebih baik. Semoga skripsi ini juga bermanfaat bagi pembaca dalam menambah wawasan serta ilmu terkait analisis kluster terutama algoritma CLARANS.

Bandung, Juli 2021

Penulis

ABSTRAK

Analisis kluster merupakan salah satu teknik multivariat yang mempunyai tujuan utama yaitu mengelompokkan objek – objek pengamatan ke dalam beberapa kluster/kelompok berdasarkan karakteristik yang dimilikinya, sehingga objek – objek yang berada di kluster yang sama memiliki tingkat kemiripan yang lebih tinggi dibandingkan dengan objek yang berada di kluster yang berbeda. Secara umum, metode analisis kluster dibagi menjadi dua yaitu metode hierarki dan partisi. Metode partisi dapat dibagi menjadi dua yaitu *k-means* dan *k-medoid* (PAM/CLARA). Kelemahan metode - metode tersebut yaitu kurang efektif dan efisien dalam melakukan analisis kluster pada dataset besar. Untuk mengatasi permasalahan ini maka digunakanlah algoritma CLARANS (*Clustering Large Application based on RANdomized Search*). CLARANS merupakan algoritma yang mengkombinasikan teknik *sampling* dan algoritma PAM. Pada studi kasus tingkat kemiskinan di 221 kota/kabupaten di Indonesia tahun 2020, algoritma CLARANS terbukti efektif dan efisien dalam melakukan analisis kluster pada dataset besar. Hal ini terlihat dari kualitas hasil klusterisasi yang cukup baik dengan nilai *silhouette coefficient* sebesar 0,6009 untuk jumlah kluster sebanyak 4 kluster. Setiap kluster memiliki jumlah anggota dan pusat kluster yang berbeda - beda dengan total *cost* sebesar 366,4794. Karakteristik yang terbentuk dari setiap kluster yaitu kluster 1 memiliki tingkat kemiskinan tertinggi, kluster 2 memiliki tingkat kemiskinan cukup tinggi, kluster 3 dan 4 masing - masing memiliki tingkat kemiskinan sedang dan rendah. Karena kluster 1 didominasi oleh kota/kabupaten di provinsi Nusa Tenggara Timur, maka kota/kabupaten di provinsi tersebut memerlukan perhatian yang lebih dibandingkan dengan kota/kabupaten di provinsi yang lain.

Kata Kunci : Analisis Kluster, *Sampling*, PAM, CLARANS, Kemiskinan.

ABSTRACT

Cluster analysis is one of the multivariate techniques that has the main goal of grouping objects of observation into several clusters/groups based on their characteristics, so that objects in the same cluster have a higher similarity than objects in different cluster. Cluster analysis method are divided into two, hierarchical and partition. Partition method are *k-means* and *k-medoids* (PAM/CLARA). The Weakness of these algorithms are the ineffective and efficient in performing cluster analysis on large datasets. Therefore, to solve this problem the CLARANS (*Clustering Large Application based on RANdomized Search*) is used. CLARANS is an algorithm that combine sampling technique and PAM algorithm. In a case study of poverty rates in 221 cities/districts in Indonesia on 2020, the CLARANS algorithm has proven to be effective and efficient in performing cluster analysis on large datasets. This can be seen from the quality of the clustering results which are good enough with a *silhouette coefficient* value 0,6009 for 4 clusters. Each cluster has a different number of members and medoids with total cost 366,4794. The characteristics formed from each cluster are that cluster 1 has the highest proverty rate, cluster 2 has a rather high poverty rate, clusters 3 and 4 each have moderate and low poverty rates. Because cluster 1 is dominated by cities/districts in the Province of East Nusa Tenggara, so the cities/districts in that province require more attention than cities/districts in other provinces.

Keywords : Cluster Analysis, Sampling, PAM, CLARANS, Poverty

DAFTAR ISI

LEMBAR HAK CIPTA.....	i
LEMBAR PENGESAHAN	ii
LEMBAR PERNYATAAN.....	iii
KATA PENGANTAR	iv
ABSTRAK.....	vi
ABSTRACT.....	vii
DAFTAR ISI.....	viii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN.....	xiii
BAB I.....	1
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	5
1.3 Tujuan Penelitian	5
1.4 Pembatasan Masalah.....	5
1.5 Manfaat Penelitian	6
BAB II.....	7
LANDASAN TEORI.....	7
2.1 Data Mining	7
2.2 Analisis Multivariat	9
2.2.1 Pengertian Analisis Multivariat	9
2.2.2 Vektor <i>Mean</i> dan Matriks Varian-Kovarian Data Multivariat.....	10
2.2.3 Nilai Eigen dan Vektor Eigen	11
2.3 Analisis Kluster.....	12

2.3.1 Pengertian Analisis Kluster	12
2.3.2 <i>Dissimilarities</i> dan <i>Similarities</i>	14
2.3.3 Asumsi dalam Analisis Kluster	17
2.3.4 Metode Analisis Kluster	20
2.3.5 Evaluasi Klasterisasi	23
2.4 Kemiskinan	27
2.4.1 Definisi Kemiskinan dan Garis Kemiskinan	27
2.4.2 Faktor Penyebab Kemiskinan	29
2.4.3 Dampak yang Ditimbulkan dari Kemiskinan	30
2.4.4 Indikator Penilaian Tingkat Kemiskinan	31
2.4.5 Upaya Pemerintah Menanggulangi Kemiskinan Akibat Pandemi Covid-19	31
BAB III	33
METODE PENELITIAN	33
3.1 Sumber Data dan Variabel Penelitian	33
3.2 Metode Analisis Data	34
3.2.1 Partitioning Around Medoids (PAM)	36
3.2.2 Clustering Large Application based on RANdomized (CLARANS)	37
3.2.3 Simulasi Penerapan Algoritma CLARANS	41
3.3 Tahapan Penelitian	46
3.3.1 Prapemrosesan Data	47
3.3.2 Melakukan Pengujian Asumsi	47
3.3.3 Menentukan Banyaknya Kluster	48
3.3.4 Menerapkan Algoritma CLARANS	48
3.3.5 Interpretasi Hasil Klasterisasi	48
3.3.6 Validasi dan <i>Profiling</i>	48
BAB IV	49
PEMBAHASAN	49
4.1 Analisis Deskriptif	49

4.2 Pengujian Asumsi	50
4.2.1 Uji Keberadaan Pencilan.....	50
4.2.2 Uji Multikolinearitas	51
4.3 Penerapan Analisis Kluster	53
4.3.1 Perancangan Program R untuk Algoritma CLARANS.....	53
4.3.2 Penentuan Banyaknya Kluster	58
4.3.3 Penentuan Jumlah <i>Maxneighbor</i> dan <i>Numlocal</i>	58
4.3.4 Hasil Analisis Kluster dengan CLARANS	61
4.3.5 Validasi Hasil Klasterisasi	66
BAB V	67
PENUTUP	67
5.1 Kesimpulan	67
5.2 Saran	68
DAFTAR PUSTAKA	69
LAMPIRAN.....	73

DAFTAR TABEL

Tabel 2. 1 Data Multivariat.....	10
Tabel 2. 2 Metode – Metode Analisis Kluster	20
Tabel 2. 3 Kriteria <i>Silhouette Coefficient</i>	27
Tabel 3. 1 Daftar Nama 16 Provinsi yang Digunakan dalam Penelitian.....	33
Tabel 3. 2 Nilai Matematika dan Fisika Siswa	41
Tabel 3. 3 Penempatan Objek <i>Nonmedoid</i> B, C, dan D.....	42
Tabel 3. 4 Penempatan Objek <i>Nonmedoid</i> B, C, dan E	43
Tabel 3. 5 <i>Cost Node</i> (A,E) dan (A,D).....	44
Tabel 3. 6 <i>Cost Node</i> (A,D) dan (A,C)	44
Tabel 3. 7 Penempatan Objek <i>Nonmedoid</i> A, C, dan E.....	45
Tabel 3. 8 <i>Cost Node</i> (A,D) dan (B,D)	45
Tabel 4. 1 Statistik Deskriptif Data Kemiskinan di Indonesia Tahun 2020.....	49
Tabel 4. 2 Nama Objek yang Merupakan Pencilan.....	50
Tabel 4. 3 Matriks Korelasi	51
Tabel 4. 4 Total Varian	52
Tabel 4. 5 Komponen Utama	53
Tabel 4. 6 Data Masukan	53
Tabel 4. 7 Data Keluaran	54
Tabel 4. 8 Statistik Deskriptif Klaster 1.....	62
Tabel 4. 9 Statistik Deskriptif Klaster 2.....	63
Tabel 4. 10 Statistik Deskriptif Klaster 3.....	64
Tabel 4. 11 Statistik Deskriptif Klaster 4.....	64
Tabel 4. 12 Karakteristik Setiap Klaster	65

DAFTAR GAMBAR

Gambar 3. 1 Proses Algoritma CLARANS	40
Gambar 3. 2 Hasil Akhir Klasterisasi	46
Gambar 3. 3 Tahapan Penelitian	47
Gambar 4. 1 Tampilan Utama Aplikasi RStudio	54
Gambar 4. 2 Cara Menambahkan R <i>Script</i>	55
Gambar 4. 3 Kode Algoritma CLARANS pada Program R	56
Gambar 4. 4 Cara Menjalankan Program CLARANS pada RStudio	56
Gambar 4. 5 Memilih Format <i>Output</i>	57
Gambar 4. 6 Hasil Metode Elbow	58
Gambar 4. 7 Grafik Pengaruh <i>Maxneighbor</i> dan <i>Numlocal</i> terhadap Kualitas Kluster	59
Gambar 4. 8 Grafik Pengaruh <i>Maxneighbor</i> dan <i>Numlocal</i> terhadap <i>Runtime</i>	60

DAFTAR LAMPIRAN

Lampiran 1. Data Tingkat Kemiskinan di 221 Kota/Kabupaten di Indonesia Tahun 2020...	73
Lampiran 2. Jarak Mahalanobis Setiap Objek	80
Lampiran 3. Gugus Data Baru Hasil Analisis Komponen Utama	83
Lampiran 4. Perbandingan Nilai Rata-Rata <i>Silhouette Coefficient</i> saat k=3 dan k=4	90
Lampiran 5. Hasil Analisis Kluster	91

DAFTAR PUSTAKA

- Amalia, R. (2007). *Analisis Deteksi Outlier Menggunakan CLARANS*. Retrieved Februari 21, 2021, from Telkom University: <https://repository.telkomuniversity.ac.id/pustaka/94296/analisis-deteksi-outlier-menggunakan-clarans-clustering-large-application-based-on-randomized-search-.html>
- Anton, H., & Rorres, C. (2004). *Aljabar Linear Elementer Edisi Kedelapan Jilid 1*. Jakarta: Erlangga.
- Badan Pusat Statistik. (2021, Januari 1). *Kemiskinan dan Ketimpangan*. Retrieved April 14, 2021, from Badan Pusat Statistik: <https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html>
- Badan Pusat Statistik. (2021, Februari 15). *Persentase Penduduk Miskin September 2020 naik menjadi 10,19 persen*. Retrieved April 14, 2021, from Badan Pusat Statistik: <https://www.bps.go.id/pressrelease/2021/02/15/1851/persentase-penduduk-miskin-september-2020-naik-menjadi-10-19-persen.html>
- Bappenas. (2004, Juni 11). *Strategi Nasional Penanggulangan Kemiskinan*. Retrieved April 14, 2021, from Badan Perencanaan dan Pembangunan Nasional: https://www.bappenas.go.id/index.php/download_file/view/7314/1200/
- Binty, T. (2019, Juli 8). *K-Medoids/Partitioning Around Medoids (PAM) – Non Hierarchical Clustering with R*. Retrieved Mei 8, 2021, from Medium.com: <https://medium.com/@tribinty/k-medoids-partitioning-around-medoids-pam-non-hierarchical-clustering-with-r-9d0af590bbc0>
- Bowo, K., Hoyyi, A., & Mukid, M. (2013). Analisis Faktor-Faktor yang Mempengaruhi Keputusan Pembelian dan Kepuasan Konsumen pada Notebook Merek Acer (Studi Kasus Mahasiswa Universitas Diponegoro). *Jurnal Gaussian*, 29-38.
- Cahyat, A. (2007). *Mengkaji Kemiskinan dan Kesejahteraan Keluarga*. Bogor: Cifor.

- Dewi, D. A., & Pramita, D. A. (2019). Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *Jurnal Manajemen Teknologi dan Informatika*, 102-109.
- Everitt, B. S. (1980). *Cluster Analysis Second Edition*. London: Heineman Educational Books Ltd.
- Fayyad, U., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*. Retrieved from AI Magazine: <http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230>
- Gantz, J., & Reinsel, D. (2012). *THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*.
- Ghozali, I. (2016). *Aplikasi Analisis Multivariate dengan Program IBM SPSS 25 Edisi 9*. Semarang: Badan Penerbit Universitas Diponegoro.
- Gio, P. U., & Irawan, D. E. (2016). *Belajar Statistika dengan R*. Medan: USU Press.
- Gujarati, D. N., & Porter, D. C. (2012). *Dasar - Dasar Ekonometrika Edisi 5 Jilid 1*. Jakarta: Salemba Empat.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques Third*. Elsevier.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis Sixth Edition*. USA: Pearson Education, Inc.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An introduction to cluster analysis*. New Jersey: Wiley Publishers.
- Kemkes RI. (2020). *Pedoman Pencegahan dan Pengendalian Coronavirus Disease (COVID-19)*. Jakarta: Kementerian Kesehatan RI.
- Komite Penanganan Covid-19 dan Pemulihan Ekonomi Nasional. (2021, Februari 1). *Data Sebaran*. Retrieved Februari 1, 2021, from covid19: covid19.go.id
- Kompas.com. (2020, September 1). *Upaya Pemerintah Tanggulangi Kemiskinan akibat Pandemi Covid-19, dari Bansos hingga Subsidi bagi UMKM*.

Retrieved April 17, 2021, from Kompas.com:
<https://amp.kompas.com/nasional/read/2020/09/01/19424371/upaya-pemerintah-tanggulangi-kemiskinan-akibat-pandemi-covid-19-dari-bansos>

- Kuncoro, M. (2003). *Metode Riset Untuk Bisnis dan Ekonomi*. Jakarta: Erlangga.
- Muningsih, E. (2017). Optimasi Jumlah Cluster K-Means dengan Metode Elbow untuk Pemetaan Pelanggan. *Prosiding Seminar Nasional Elinvo* (pp. 105-114). Yogyakarta: Academia.
- Ng, R. T., & Han, J. (2002). CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE Transactions on Knowledge and Data Engineering* .
- Salitonga, P. (2016). Analisis Profil Mahasiswa Politeknik Negeri Batam dengan Teknik Data Mining Asosiasi dan Clustering. *Jurnal Integrasi*, 8(1), hal. 16-21.
- Siagian, M. (2012). *Kemiskinan dan Solusi*. Medan: Grasindo Monoratama.
- Sihombing, R. E., Rachmatin, D., & Dahlan, J. A. (2019). Program Aplikasi Bahasa R untuk Pengelompokan Objek Menggunakan Metode K-Medoids Clustering. *Jurnal EurekaMatika*, Vol.7, No.1.
- Suyanto. (2019). *Data Mining untuk Klasifikasi dan Klasterisasi Data Edisi Revisi*. Bandung: Informatika.
- Turner, V. (2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. Retrieved from <https://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm>.