ALGORITMA *PRINCIPAL COMPONENT ANALYSIS* UNTUK MENINGKATKAN PERFORMA *FUZZY C-MEANS* PADA KLASTERISASI DATASET BERDIMENSI

[1]Agung Riyadi, [2]Fauziah
[1,2]Magister Teknologi Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional
Jl. Sawo Manila, Pejaten, Pasar Minggu
Jakarta Selatan, DKI Jakarta 12520
[1]2022.agung.riyadi@student.unas.ac.id, [2]fauziah@civitas.unas.ac.id

Abstrak

Data berdimensi tinggi sangat sulit untuk dikelompokkan, karena pertumbuhan datanya yang eksponensial dari segi format data dan jumlah nilai setiap dimensi yang tidak mungkin untuk dihitung. Untuk meningkatkan efisiensi dan akurasi pemrosesan data berdimensi tinggi, sebelum pengujian dilakukan pembersihan data dan proses reduksi data dengan metode *Principal Component Analysis* (PCA). Terdapat tantangan dalam membandingkan kualitas klaster pada nilai keanggotaan yang berbeda sehingga menghasilkan klaster akhir yang tidak sama, hal ini disebabkan adanya *noise* dan *outlier* pada data yang diproses. Oleh karena itu metode PCA dapat digunakan untuk mereduksi kumpulan data dari data berdimensi tinggi menjadi data berdimensi rendah serta menghilangkan *noise* dan *outlier*. Metode *Fuzzy C-Means* (FCM) dengan insialisasi berbeda digunakan untuk mengelompokkan data menjadi klaster berdasarkan data yang mirip, sehingga penempatan data yang saling berkaitan di klaster yang sama. Berdasarkan proses tersebut didapatkan perbandingan hasil metode tanpa PCA dan dengan PCA dan diperoleh hasil implementasi PCA+FCM dengan inisialisasi distribusi *Gaussian multi-variate* lebih tinggi dengan tingkat akurasi 87,07%.

***Kata Kunci**: distribusi gaussian multi-variate, FCM, klaster, PCA*

Abstract

High-dimensional data is very difficult to group, because the data growth is exponential in terms of data format and the number of values for each dimension is impossible to calculate. To increase the efficiency and accuracy of processing high-dimensional data, before testing, data cleaning and data reduction processes were carried out using the Principal Component Analysis (PCA) method. There are challenges in comparing the quality of clusters at different membership values resulting in different final clusters, this is due to the presence of noise and outliers in the processed data. Therefore, the PCA method can be used to reduce data sets from high-dimensional data to low-dimensional data and eliminate noise and outliers. The Fuzzy C-Means (FCM) method with different initialization is used to group data into clusters based on similar data, so that related data is placed in the same cluster. Based on this process, a comparison of the results of the method without PCA and with PCA was obtained and the results obtained from the implementation of PCA+FCM with the initialization of a multi-variate Gaussian distribution were higher with an accuracy level of 87.07.

***Keywords**: clustering, multi-variate gaussian distribution, FCM, PCA*

PENDAHULUAN

Di banyak negara didunia, kekerasan terhadap anak bukanlah *issue* baru yang terjadi. Kekerasan merupakan tindakan yang ditujukan untuk menyakiti orang lain, tindakan ini dilakukan berulang atau dipelajari melalui pengulangan. Tindakan yang lebih cenderung

pada agresi ini memiliki bentuk diantaranya agresi fisik seperti meninju, menendang, dan lainnya. Selain itu ada agresi verbal seperti mengejek, *bullying*, berteriak, dan lainnya [1]. Sehingga dapat didefinisikan bahwa kekerasan adalah tindakan seseorang yang menyebabkan luka fisik dan psikis [2]. Berdasarkan Undang-Undang Nomor 23 Tahun 2002 Indonesia, seseorang dikategorikan anak yaitu belum mencapai umur 18 tahun, bayi didalam kandungan termasuk didalamnya [3]. Kekerasan terhadap anak memiliki kategori bentuk yaitu fisik, mental, melukai, pelecehan, penelantaran, dan eksploitasi [4]. Umumnya kekerasan anak terjadi terhadap anak yang *introvert*, sebagian besar pelaku adalah orang tua, saudara, paman, bibi, teman, guru, tetangga, dan orang asing. Dalam banyak kasus kekerasan terhadap anak tidak terjadi secara spontan tetapi juga dipengaruhi oleh keluarga atau lingkungan sekitar [5]. Sementara itu, Richard Gelles berpendapat bahwa kekerasan terhadap anak dapat dipicu melalui empat kategori seperti kekerasan warisan (belajar kekerasan dari orang tua); stres sosial (terkait dengan kondisi sosial seperti kondisi ekonomi); isolasi sosial (terjadi pada orang tua yang tidak pernah mengikuti kegiatan sosial); Struktur keluarga (tidak memiliki hubungan yang baik di dalam keluarga itu sendiri) [6].

Data yang dirilis oleh Komisi Perlindungan Anak Indonesia (KPAI) pada tahun 2022, sejumlah 4.124 aduan kasus pada rentang bulan Januari sampai November tahun 2022. Dari jumlah aduan tersebut, sebanyak 1.706 kasus hak pemenuhan berasal dari klaster keluarga dan pengasuhan alternatif, 376 klaster pendidikan, pemanfaatan waktu luang, dan kegiatan budaya dan agama. Selain hal tersebut, terdapat 1.903 aduan perlindungan khusus anak yang terbagi dalam beberapa klaster, yaitu 746 kasus korban kejahatan seksual, 454 korban kekerasan fisik-psikis, 187 anak sebagai pelaku, eksploitasi ekonomi sejumlah 80 kasus, serta 70 kasus anak korban pornografi dan kejahatan siber [7]. Tingginya kasus kekerasan pada anak yang terjadi di Indonesia, mencerminkan permasalahan kekerasan belum tertangani secara merata. Pengelompokan daerah rawan kekerasan menggunakan teknik *clustering* dapat digunakan untuk melihat pola kejadian tindak kekerasan [8].

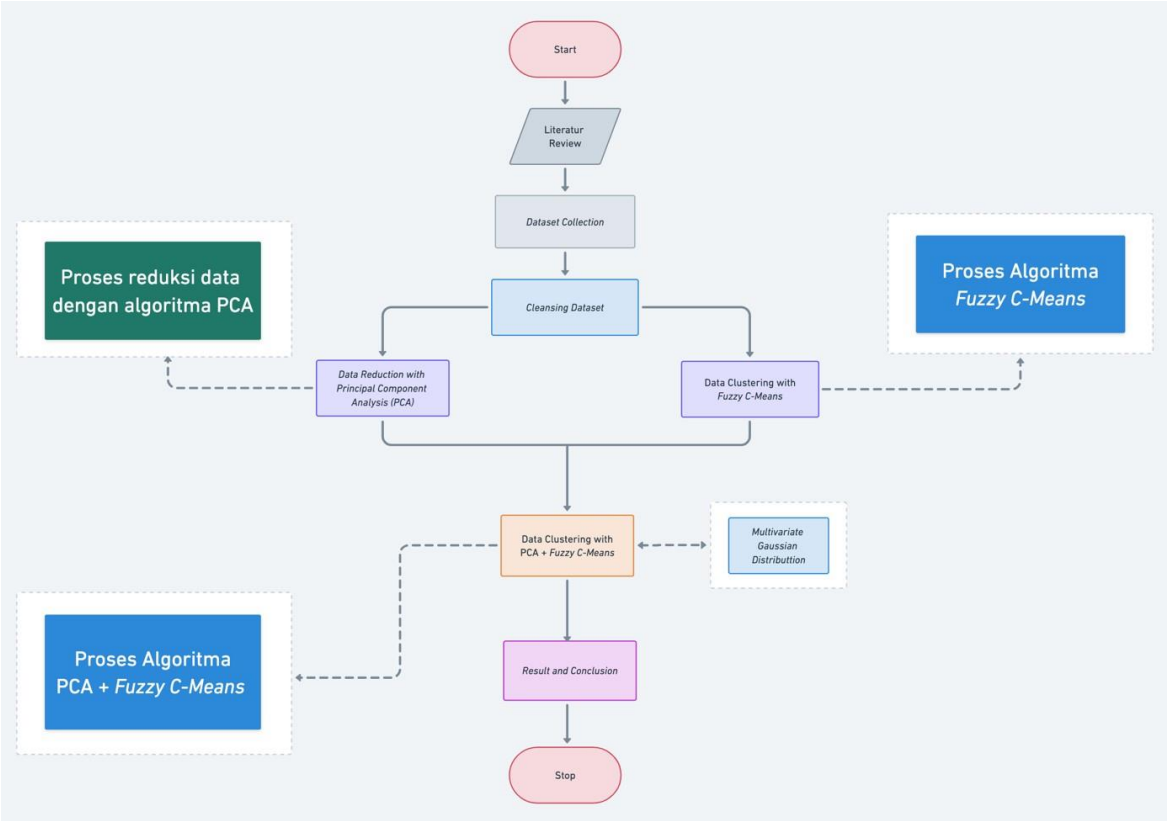
Banyak penelitian, analisis dan pengelompokan yang tepat pada kasus kekerasan terhadap anak dengan kategori tingkat kekerasan ke dalam klaster menggunakan K-Means dan *Latent Dirichlet Allocation* (LDA), memudahkan untuk mengetahui klaster wilayah rawan kekerasan, penyebab tindak kekerasan, bentuk kekerasan, dan lainnya. Dengan menganalisa dokumen kasus dan database kasus menggunakan metode *clustering*, dapat memunculkan informasi penting sehingga dapat dikelompokan dalam klaster-klaster yang telah ditentukan. Klaster yang dihasilkan dapat ditindaklanjut oleh pemerintah, pekerja sosial, dan pihak berwenang lainnya dalam menentukan langkah

dalam identifikasi dan menyusun rencana solusi dalam menangani tindak kekerasan terhadap anak [9][10][11][12]. Metode *Fuzzy C-Means* digunakan karena lebih fleksibel dalam menentukan pusat klaster dan memungkinkan poin data dalam dataset menjadi anggota 2 klaster atau lebih, tetapi terdapat kelemahan kelompok data yang tidak konsisten [13]. Algoritma *Principal Component Analysis* mereduksi dataset dengan mengganti atau menghilangkan *noise dan outlier*, sehingga dapat meningkatkan akurasi dan menjadikan dataset lebih konsisten [14].

Sumber data yang dianalisis seringkali memuat data akhir yang berasal umumnya dari database, dan mengabaikan data dokumen yang berasal dari narasi kronologi kasus serta hasil *survey* langsung. Analisis terhadap berbagai sumber data (numerik, kategorikal, dan narasi) dengan karakteristik berbeda-beda sulit dilakukan sehingga dapat mempengaruhi tingkat akurasi dan hasil pengelompokan data. Pengelompokan data menggunakan *Fuzzy C-Means* dilakukan terhadap dataset yang memiliki data memungkinkan menjadi anggota lebih dari 1 klaster, dan implementasi algoritma *Principal Component Analysis* untuk mengatasi kelemahan *Fuzzy C-Means* dalam hal inkosistensi keanggotaan kelompok data.

METODE PENELITIAN

Metodologi dalam penelitian ini dibagi dalam beberapa tahap yang disusun secara sistematik seperti dibawah ini:



Gambar 1 Usulan Metode Penelitian

Tingginya kasus kekerasan terhadap anak di Indonesia dan dinamika implementasi sistem perlindungan anak yang terkendala beberapa aspek, diantaranya hukum dan budaya serta algoritma *clustering* yang terkendala data berdimensi tinggi, serta data yang bersifat campuran kategorikal dan numerikal merupakan latar belakang pelaksanaan penelitian. Pada gambar 1 merupakan tahapan didalam melakukan penelitian, setiap tahapan memiliki proses yang pada inti prosesnya tergambar dengan hubungan garis putus-putus terdapat penghitungan algoritma PCA, *Fuzzy C-Means* dan penghitungan fungsi kepadatan peluang (fkp) dengan *Multivariate Gaussian Distribution*. Detil setiap proses dijelaskan sebagai berikut:

Studi Literatur

Tahap pemahaman teori dan pendalaman metode atau algoritma yang digunakan dalam penelitian melalui berbagai sumber referensi.

Pengumpulan Dataset

Dataset dalam penelitian ini mengambil dari data kasus kekerasan terhadap anak di Komisi Nasional Perempuan Indonesia selama tahun 2018 sampai 2022. Dataset memiliki karakteristik kategorikal dan *numerical* yang mencakup demografi entitas orang, umur, jenjang pendidikan, kronologi kejadian, dan lainnya. Proses pengumpulan data melingkupi *collecting* data bersumber dari database dan dokumen kasus, pemilihan data menjadi dataset, penggabungan data, *filtering* dan pembersihan data. Untuk kebutuhan dalam proses pengelompokan data, maka dilakukan penyederhaan struktur data yang dilakukan dan dimasukkan ke dalam format tabular berupa tabel. Data kategorikal diubah menjadi format tertentu sehingga dapat diproses dalam pengujian pengelompokan data.

Desain Algoritma

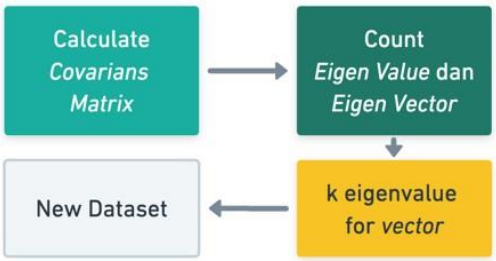
Tahap pembuatan bagan alur algoritma terhadap gambaran logika untuk memahami proses yang dilakukan dengan algoritma sebagai berikut dibawah ini:

1. *Principal Component Analysis* (PCA)

   Algoritma PCA mereduksi data dengan mentransformasi data ke dalam variabel baru (komponen utama) yang saling tidak berkorelasi untuk membentuk koordinat baru dengan variansi maksimum tetapi tidak menghilangkan karakteristik data tersebut [14][15][16][17]. Dalam penelitian ini, PCA digunakan untuk mereduksi data berdimensi tinggi sehingga hasilnya bermanfaat dalam proses pengelompokan data.

   Tahapan pada gambar 2 dalam reduksi data menggunakan PCA sebagai berikut:

   1. Dataset terstruktur dihitung *covarians matrix* untuk fitur-fitur yang ada dalam dataset
   2. Menghitung *eigen value* dan *eigen vector* untuk *covarians matrix*



Gambar 2 Bagan alur proses reduksi data menggunakan PCA

3. Mengurutkan *eigen value* dan *eigen vector* yang sesuai kemudian memilih k nilai eigen (*eigen value*) untuk membentuk matriks *eigen vector*
4. Mengubah ke matriks asli sehingga didapatkan dataset baru hasil proses reduksi data

2. Desain *flow chart Fuzzy C-Means*

Clustering merupakan proses untuk mengelompokan data menjadi klaster berdasarkan data yang mirip, sehingga penempatan data yang saling berkaitan di *cluster* yang sama [18].

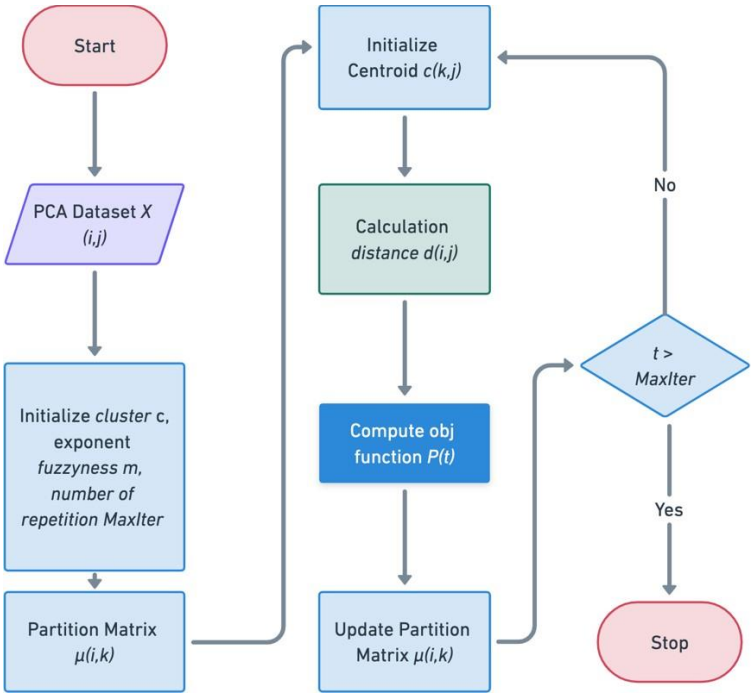
Metode *fuzzy* sejak diperkenalkan oleh Zadeh, diimplementasikan dalam algoritma clustering secara cepat. *Fuzzy C-Means* (FCM) salah satu algoritma paling terkenal dan memperoleh hasil pengelompokan dengan meminimalkan fungsi tujuan dan iterasi keanggotaan dan *centroid* [19]. Fungsi FCM secara umum seperti pada persamaan 2.1 dibawah ini:

$$f_{fcm}(a,b) = \sum_{i=1}^{c}\sum_{f=1}^{n} a_{if}^{m}\, D_{if}^{2}, m > 1 \quad (2.1)$$

Algoritma *Fuzzy* menunjukkan performa lebih baik dengan jumlah beberapa perulangan, namun hal tersebut membuat algoritma *Fuzzy* menjadi *sensitive* terhadap *noise* pada data dan biasanya mengarah ke titik deviasi pusat untuk titik data anomali individual [20]. Algoritma *Fuzzy* di-tingkatkan dan menambahkan algoritma C-Means yang dapat melonggarkan constraints [21].

Pengelompokan data dengan setiap titik memiliki level keanggotaan masing-masing dan dikelompokan ke dalam pusat *cluster* data. Alur Algoritma *Fuzzy C-Means* dalam implementasi pada dataset sebagai berikut dibawah ini.

Metode FCM memulai dengan menetapkan pusat klaster untuk merepresentasikan lokasi rata-rata dari setiap klaster. Setiap titik data diberi derajat keanggotaan untuk setiap klaster.



Gambar 3 Alur Algoritma *Fuzzy C-Means*

Dengan mengoreksi pusat klaster dan derajat keanggotaan titik data secara berulang, pusat klaster akan bergerak menuju posisi yang tepat. Perulangan dilakukan berdasarkan derajat keanggotaan, yang mencerminkan jarak titik data dari pusat klaster yang memiliki bobot sesuai derajat keanggotaan titik data tersebut. [22]. Dan pada beberapa tahun terakhir, banyak jarak matrik diusulkan untuk meningkatkan kinerja algoritma *Fuzzy C-Means* [23].

Pada gambar 3, alur algoritma *fuzzy C-Means* dijelaskan dengan runutan sebagai berikut:

1. Dataset PCA, data berdimensi atau ruang fitur yang telah direduksi melalui proses PCA
2. Inisialisasi jumlah *cluster*, eksponen, dan jumlah perulangan
3. Penetapan matriks partisi *fuzzy* pada rentang 0 hingga 1
4. Penghitungan pusat *cluster* dengan persamaan seperti dibawah ini

$$v_{jk} = \frac{\sum_{i=1}^{n} \mu_{ik}^{m} x_{ij}}{\sum_{i=1}^{n} \mu_{ik}^{m}} \qquad (2.2)$$

5. Menghitung jarak menggunakan *euclidean distance*
6. Menghitung fungsi tujuan *P(t)*
7. Memperbarui matriks partisi *fuzzy*

Iterasi tahap 4 – 7 sampai kriteria berhenti terpenuhi

HASIL DAN PEMBAHASAN

Dataset penelitian didapatkan dari Lembaga aduan kasus anak sejumlah 150 data

dengan 21 atribut. Implementasi penelitian menggunakan Bahasa pemrograman *Python* dengan Pustaka *Science-Kit Learn* (sklearn) dan *Science-Kit Fuzzy* (skfuzzy), penerapan algoritma PCA dan FCM dilakukan didalam kode program Python.

Tabel 1. Dataset Demografi Kasus

ranah_kasus	**media_pengaduan**	**pendidikan_korban**	**pekerjaan_korban**	**usia_korban**	**status_pernikahan_korban**
Relasi Personal	Google Forms	SMA	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Surat Elektronik	SMA	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Google Forms	SMA	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Google Forms	SMK	-	18	Belum Kawin
Relasi Personal	Google Forms	SMK	Wirausaha	18	Belum Kawin
Relasi Personal	Datang langsung	SMA	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Surat Elektronik	SMP	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Telephone	SMP	PRT (Pekerja Rumah Tangga)	18	Belum Kawin
Relasi Personal	Datang langsung	SMA	Tidak Bekerja	18	Belum Kawin
Relasi Personal	Google Forms	SMP	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Google Forms	SMA	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Telephone	SMA	Penggiat HAM/Pendamping/LSM	18	Belum Kawin
Relasi Personal	Surat Elektronik	-	-	18	-

Relasi Personal	Telephone	SMA	Pelajar/Mahasiswa	18	Kawin Tidak Tercatat
Relasi Personal	Surat Elektronik	SMA	Ibu Rumah Tangga	18	Kawin Tidak Tercatat
Relasi Personal	Google Forms	SMP	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Datang langsung	SMP	PRT (Pekerja Rumah Tangga)	18	Belum Kawin
Relasi Personal	Pengaduan	SMK	Pegawai Swasta	18	Belum Kawin
Relasi Personal	Google Forms	SMP	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Google Forms	SMP	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Google Forms	SMP	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Google Forms	SMA	Pelajar/Mahasiswa	18	Belum Kawin
Relasi Personal	Google Forms	SMA	Pelajar/Mahasiswa	18	Belum Kawin

***Cleansing* Data**

Data didalam dataset berupa teks yang perlu dilakukan *cleansing* karena didalamnya terdapat data dengan tipe *string,* agar data menjadi memiliki tipe yang konsisten sebelum dilakukan pengolahan. Untuk memudahkan proses dalam implementasi PCA, dilakukan konversi string pada beberapa isian atribut dari teks menjadi angka.

Gambar 4 merupakan bagian dari dataset yang ditampilkan 5 baris pertama dari keseluruhan data dan sebelum dilakukan proses konversi label atau format string menjadi angka.

Hasil konversi menggunakan *Label-Encoder* pada Pustaka *sklearn* di Bahasa pemrograman Python direpresentasikan pada gambar 5.

tanggal_terjadi	waktu_terjadi	jenis_pelaku	lokasi_kejadian	provinsi	kabupaten_kota	...	bentuk_kekerasan	intensitas_kekerasan	modus_kekerasan
2021	Lainnya	Individu	NaN	DKI JAKARTA	KOTA ADM. JAKARTA BARAT	...	Psikis : - Diancam	Lebih dari sekali	Kekerabatan
2018	-	Individu	Kota Bandung, Jawa Barat.	JAWA BARAT	KOTA BANDUNG	...	Fisik : - Pemukulan; \nFisik : - Ditampar	Lebih dari sekali	Kekerabatan
2022	Lainnya	Individu	Di tempat korban (melalui media)	JAWA TIMUR	BANYUWANGI	...	Seksual : - Ancaman Penyebaran Foto/Vidio Porno	Lebih dari sekali	Kekerabatan
2021	Lainnya	Individu	NaN	KALIMANTAN SELATAN	KOTA BANJARMASIN	...	Seksual : - Ancaman Penyebaran Foto/Vidio Porno	Sekali	Kekerabatan
2021	Lainnya	Individu	Tidak teridentifikasi.	DKI JAKARTA	KOTA ADM. JAKARTA TIMUR	...	Seksual : - Ancaman Penyebaran Foto/Vidio Porno	Lebih dari sekali	Kekerabatan

Gambar 4 Dataset sebelum dilakukan proses konversi string

	Id	kategori_kasus	tanggal_pengaduan	tanggal_terjadi	waktu_terjadi	jenis_pelaku	lokasi_kejadian	provinsi	kabupaten_kota	konteks_kekerasan	..
0	1	0	83	18	1	1	33	4	20	2	.
1	2	0	40	16	0	1	23	7	27	6	.
2	3	0	47	19	1	1	3	9	7	2	.
3	4	0	107	18	1	1	33	11	29	2	.
4	5	0	94	18	1	1	31	4	22	2	.

Gambar 5 Dataset setelah dilakukan konversi string menjadi angka

	ranah_kasus	jenis_kekerasan	pekerjaan_korban	status_pernikahan_korban	pendidikan_korban
0	8	14	4	1	SMA
1	6	9	4	1	SMA
2	4	11	3	1	SMA
3	3	10	5	1	SMA
4	7	15	4	1	SMA

Gambar 6 Dataset yang memiliki isian angka dalam 5 baris pertama

Proses konversi menggunakan fungsi *LabelEncoder* terhadap dataset dilakukan terhadap beberapa komponen yang menjadi variabel inti. Gambar 5 diatas merupakan hasil konversi terhadap variabel yang dipilih dari value string menjadi angka, tujuannya digunakan untuk pemrosesan pada algoritma PCA dan FCM.

Tahap *cleansing* selanjutnya untuk penyiapan data angka yang konsisten untuk pemrosesan pada tahap selanjutnya yaitu menghapus kolom-kolom yang memiliki *value* kosong atau tidak terdefinisikan, sehingga menghasilkan dataset yang hanya memiliki isian angka seperti pada gambar 6.

Gambar 6 menampilkan data 5 baris pertama dari total data sejumlah 150 data yang secara format data memiliki tipe angka setelah dilakukan normalisasi data.

Implementasi PCA untuk reduksi data

Hasil pada tahap sebelumnya telah didapatkan data yang hanya berisi angka, reduksi data yang dilakukan pada tahap implementasi PCA menggunakan pustaka sklearn. Dalam proses reduksi data dengan PCA

juga dilakukan proses cek *noise* dan *outliers* serta menghapusnya dari dataset. Penghapusan *noise* dan *outliers* bertujuan untuk mengurangi duplikasi data didalam klaster yang menyebabkan kelompok data menjadi tidak konsisten dan jarak data yang tidak jelas.

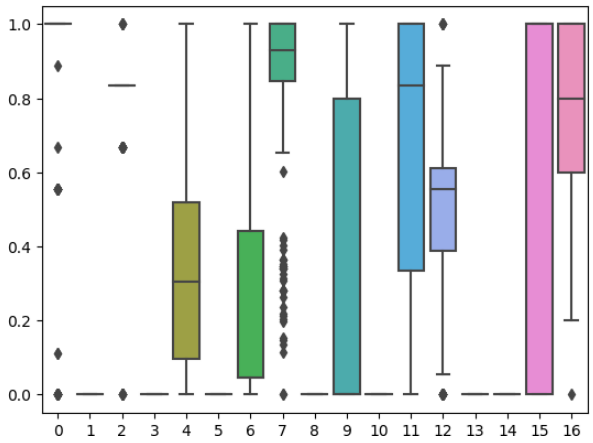
Hasil proses pengecekan dan penghapusan *outliers* seperti pada gambar 7, didapatkan 3 *index dataset* yang dihapus yang merupakan *outliers* sehingga dataset yang fit berjumlah 147. Dataset yang telah fit dinormalisasikan dengan metode transformasi *MinMaxScalar* sehingga didapatkan grafik pada gambar 8.

Grafik skalar pada gambar 8 diatas menggunakan metode normalisasi data *Min-Max* yang menghitung setiap nilai pada sebuah fitur data dengan tujuan menskalakan data ke dalam rentang tetap antara 0 dan 1, sehingga menghasilkan distribusi *mean* dan varian satuan nol.

Implementasi algoritma PCA dengan nilai komponen 2 pada dataset hasil normalisasi *MinMaxScalar* menghasilkan data pada table 2.

	ranah_kasus	jenis_kekerasan	pekerjaan_korban	status_pernikahan_korban
0	8	14	4	1
1	6	9	4	1
2	4	11	3	1
3	3	10	5	1
4	7	15	4	1

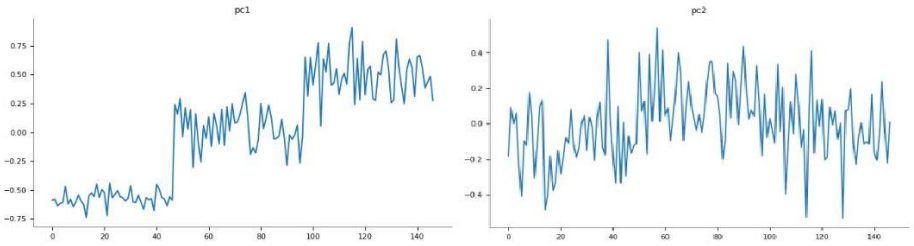
Gambar 7 Data yang dihapus berdasarkan hasil pengecekan *noise* dan *outliers*



Gambar 8 Grafik Skalar Hasil Normalisasi Data

Table 2 Dataframe dataset hasil implementasi PCA

	pc1	pc2
0	-0.587317	-0.181048
1	-0.581580	0.088521
2	-0.637457	0.000453
3	-0.616700	0.057697
4	-0.608234	-0.225287
...	...	...
142	0.555087	-0.049118
143	0.383280	0.235856
144	0.435931	-0.034722
145	0.485378	-0.219976
146	0.276550	0.007536



Gambar 9 Grafik Nilai 2 Komponen Prinsipal PCA

Nilai-nilai komponen principal (pc1 dan pc2) pada dataframe tabel 2 merupakan hasil implementasi PCA dengan nilai komponen 2.

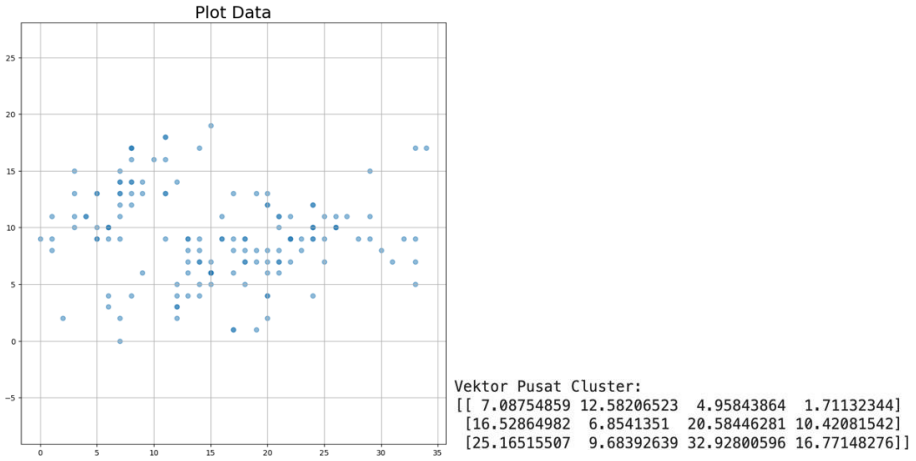
Komponen principal pada implementasi PCA berdasarkan tabel 2 diproyeksikan ke dalam grafik 2 dimensi pada gambar 9, terdapat perbedaan data pada komponen principal 1 (pc1) dan komponen principal 2 (pc2). Dimana pada pc1 terlihat data lebih bervariasi dibanding dengan pc2.

Pengelompokan menggunakan FCM

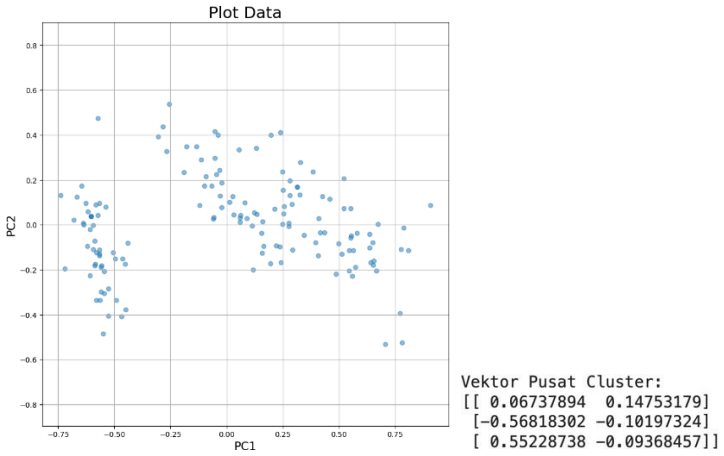
Pengelompokan data menggunakan FCM menggunakan distribusi *Gaussian multi-variate* dengan 3 klaster karena pada dataset memiliki 3 kategori jenjang pendidikan korban berdasarkan dataset awal pada tabel 1, parameter MAX_ITER sebagai batas maksimal pengulangan sebanyak 100 kali, panjang data point yaitu sebanyak dataset yang dilakukan pengujian. Detil parameter yang didefinisikan pada pengelompokan FCM sebagai berikut:

k = 3
MAX_ITER = 100
n = *panjang (dataset)*

Pengujian pertama dengan menerapkan terhadap dataset tanpa implementasi PCA:



Gambar 10 Grafik pengelompokan data menggunakan FCM tanpa PCA



Gambar 11 Grafik hasil implementasi PCA dalam FCM

Berdasarkan gambar 10, pengelompokan data menggunakan FCM tanpa PCA diketahui bahwa pada grafik terdapat *noise* dan *outlier* sehingga mempengaruhi pembentukan klaster dan *vector* pusat klaster dengan nilai akurasi 86,67%. Data *noise* dan *outlier* merupakan data duplikat, bernilai salah atau anomali dan memiliki karakteristik berbeda jauh sehingga dapat menyebabkan kelompok data menjadi tidak konsisten dan jarak data yang tidak jelas.

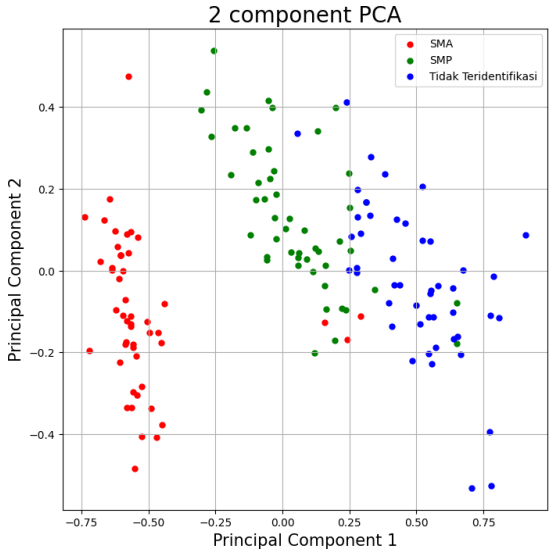
Proses PCA membantu mengatasi masalah data dimensi tinggi dengan menghapus *noise* dan *outlier* sehingga menghasilkan model FCM yang lebih efisien dan meningkatkan nilai akurasi.

Gambar 11 menunjukkan data-data point yang semakin bergabung ke dalam klaster dengan jarak yang lebih dekat setelah reduksi data dengan PCA, nilai akurasi yang didapatkan lebih baik dari pengujian pertama yang tidak mengimplementasikan PCA yaitu 87,07%. Terdapat peningkatan akurasi yang

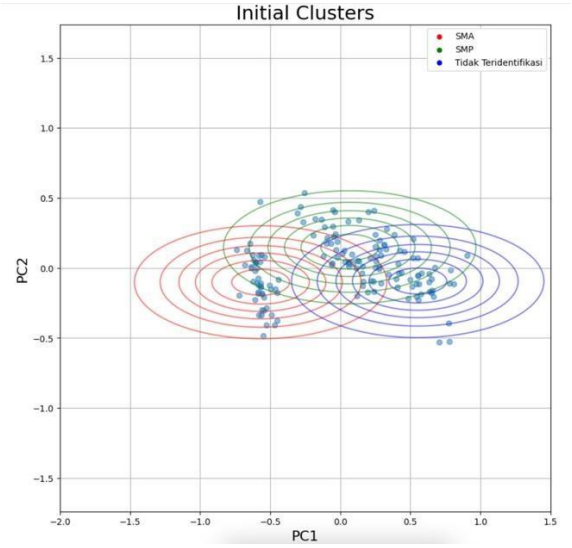
tidak signifikan antara FCM tanpa PCA dengan implementasi PCA pada FCM, hal ini kemungkinan disebabkan dataset yang didapatkan tergolong sedikit.

Klaster Ranah Kekerasan Terhadap Anak Pada Jenjang Pendidikan

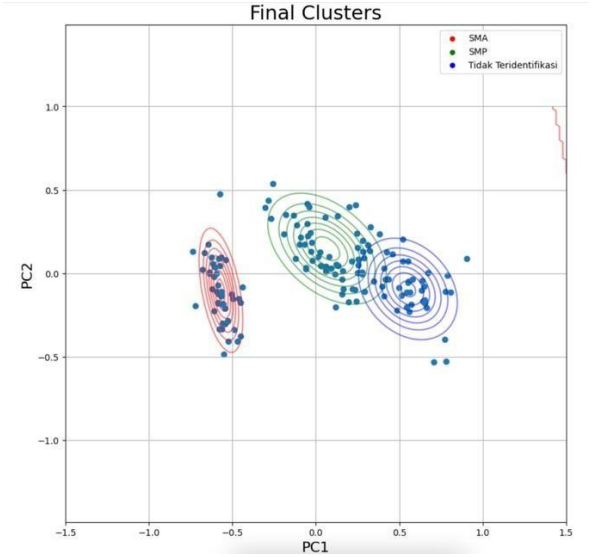
Klaster yang terbentuk dengan FCM dengan metode distribusi *Gaussian multi-variate* dan implementasi PCA didalam prosesnya terhadap dataset yang telah dinormalisasi datanya dikelompokkan berdasarkan jenjang pendidikan korban, yaitu SMA (berwarna merah), SMP (berwarna hijau) dan Tidak Teridentifikasi (berwarna biru).



Gambar 12 Informasi klaster dalam 2 komponen PCA



Gambar 13 Inisial klaster PCA+FCM



Gambar 14 Final klaster PCA+FCM

Informasi klaster ranah kekerasan terhadap anak pada jenjang pendidikan dikelompokkan seperti yang terlihat pada gambar 9 berdasarkan komponen principal 1 dan komponen principal 2 dalam perhitungan PCA dengan jumlah komponen 2. Melalui pengelompokan dengan metode FCM dan distribusi *Gaussian multi-variate* diperoleh area inisial klaster seperti pada gambar 13 yang menjelaskan area klaster data berdasarkan kelompok jenjang pendidikan.

Final klaster pada gambar 14 menunjukkan bahwa dengan pencarian mode terdekat dan penghitungan indeks data masing-masing kelompok didapatkan grafik klaster data yang semakin spesifik dengan kelompok jenjang SMA lebih terkumpul data-data point-nya dan lebih tersebar pada kelompok jenjang SMP dan Tidak Teridentifikasi. Hasil penelitian yang didapatkan mungkin karena *centroid* awal yang dihasilkan oleh algoritma PCA yang diusulkan cukup dekat dengan solusi optimal dan juga menemukan klaster di ruang berdimensi rendah untuk mengatasi permasalahan dimensi dengan cara menghilangkan *noise* dan *outliers* pada data.

KESIMPULAN DAN SARAN

Implementasi *Principal Component Analysis* (PCA) pada *Fuzzy C-Means* (FCM) *Clustering* memberikan manfaat signifikan dalam mengatasi masalah dimensi tinggi, meningkatkan efisiensi perhitungan, dan memahami variabilitas data dengan mempertahankan informasi yang paling relevan. PCA membantu menyederhanakan struktur data, meningkatkan keterbacaan hasil *clustering*, dan potensial untuk mengatasi masalah *overfitting*. Implementasi PCA dalam proses *Fuzzy C-Means* dan penggabungan Distribusi *Gaussian Multi-Variate* menunjukkan hasil akurasi 87,07% pada 150

data dengan parameter 3 klaster dan 100 kali iterasi. Namun, perlu hati-hati dalam pemilihan komponen utama agar informasi yang signifikan tidak hilang, dan evaluasi hasil *clustering* diperlukan untuk memastikan kualitas dan interpretasinya.

Saran dan rekomendasi yang dapat disampaikan bahwa analisis klasterisasi menggunakan PCA dan FCM diperoleh 3 klaster pada jenjang Pendidikan korban yaitu SMP, SMA dan Tidak Teridentifikasi. Penelitian dapat dilakukan kedepannya dengan dataset yang lebih banyak dengan dimensi data yang lebih beragam, sehingga dapat memungkinkan munculnya jenis klaster lain seperti demografi dan lainnya. Analisis sederhana terhadap dataset menggunakan BigQuery dan *Looker Studio* didalam *Google Cloud Platform* dapat dilihat jika di semua jenjang pendidikan, jumlah korban yang mengalami kekerasan seksual berbasis elektronik paling besar. Hal ini dapat menjadi *concern* bahwa kemajuan teknologi dan transformasi digital memiliki sisi negatif yang berdampak pada aktifitas tidak baik bagi penggunanya, sehingga diperlukan literasi digital yang baik dan kesadaran atas keamanan diri dan data pribadi dalam melakukan aktifitas di ranah digital.

DAFTAR PUSTAKA

[1] K. A. Nelson, R. J. Davis, D. R. Lutz, dan W. Smith, "Optical generation of tunable ultrasonic waves," *Journal of Applied Physics*, vol. 53, no. 2, Feb., hal. 1144 – 1149, 2002.

[2] Nurhayati, Nurhayati & Setyani, I. (2021). Trauma Masa Anak-Anak Dan Perilaku Agresi. Psikobuletin: Buletin Ilmiah Psikologi. 2. 164. 10.24014/pib.v2i3.13917

[3] Ferrara, P., Franceschini, G., Villani, A. et al. (2019). Physical, psychological and social impact of school violence on children. Ital J Pediatr 45, 76. https://doi.org/10.1186/s13052-019-0669-z

[4] Melati, K., & Parwata, A. (2022). Perlindungan Hukum Atas Perkawinan Anak Di Bawah Umur Dalam Perspektif Undang-Undang Hak Asasi Manusia. *Kertha Semaya: Journal Ilmu Hukum*, *10*(9), 1994-2002. doi:10.24843/KS.2022.v10.i09.p03

[5] Al-Mohannadi AS, Al-Harahsheh S, Atari S, Jilani N, Al-Hail G and Sigodo K. (2022). Addressing violence against children: A case review in the state of Qatar. Front. Public Health 10:859325. doi: 10.3389/fpubh.2022.859325

[6] Dasadwiasting, Valentia Nadya. (2022). The Dynamic of Child Protection System UNICEF to Reducing Violence Against Children in Indonesia. Indonesian Journal of Multidisciplinary Science E-ISSN: 2808-6724

[7] Rahayu, Rita, & Day, John. (2015). Determinant factors of e-commerce adoption by SMEs in developing country: evidence from Indonesia.

Procedia-Social and Behavioral Sciences, 195, 142–150

[8] KPAI Catat 4.124 Kasus Perlindungan Anak hingga November 2022. https://dataindonesia.id/ragam/detail/kpai-catat-4124-kasus-perlindungan-anak-hingga-november-2022 diakses 22 Januari 2023 Pkl. 22:45 WIB

[9] Rahma, Raisya., & Mufidah, Ratna. (2022). Pengelompokan Daerah Rawan Kekerasan Terhadap Perempuan dan Anak di Jawa Barat Menggunakan Algoritma K-Means. Jurnal Ilmiah Penelitian dan Pembelajaran Informatika (JIPI) Vol. 07 No.03, 850-857

[10] Rahmah, Yuni Shafira., & Kirana, Kartika Chandra. (2022). The Implementation of Child-Friendly City Programs in Special Protection Cluster at Serang-Banten Province. Journal Studi Gender dan Anak (JSGA) Vol. 9, No. 02

[11] Adawiyah, Noviy., Sulistiyowati, Nina., & Jajuli, Mohamad. (2021). Klasterisasi Kasus Terhadap Anak dan Perempuan Berdasarkan Algoritma K-Means. Generation Journal Vol. 5 No. 2

[12] Tresnasari, Nur Annisa., Adji, Teguh Bharata., & Permanasari, Adhistya Erna. (2020). Social-Child-Case Document Clustering based on Topic Modelling using Latent Dirichlet Allocation. Indonesia Journal of Computing and Cybernetics Systems (IJCCS) Vol. 14 No. 2, 179-188

[13] Surono, Sugiyarto., & Putri, Rizki Desia Arindra. (2021). Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis. *International journal of fuzzy systems* (1562-2479), 23 (1), p. 139.

[14] Boothby, Neil & Stark, Lindsay. (2011). Data Surveillance in Child Protection Systems Development: An Indonesian Case Study. Elsevier: Child Abuse & Neglect 35 (2011) 993-1001

[15] Annisa, Ayu. (2020). Speech Act on Conversational Argumentation: A Study of Pragmatic In Cable News Network.Google Scholar

[16] Aziz. D.Z. Sulianta F. (2022) penggunaan google cloud platform untuk marketeer dan analis dalam pengolahan data. Jurnal Syntax Idea 4 (9)

[17] Sari, Y.P., Primajaya, A., & Irawan, A. S. Y. (2020). Implementasi Algoritma K-Means untuk Clustering Penyebaran Tuberkulosis di Kabupaten Karawang. INOVTEK Polbeng - Seri Inform., vol. 5, no. 2, p. 229, doi: 10.35314/isi.v5i2.1457

[18] Chen, Jiashun., Zhang, Hao., Pi, Dechang., Kantardzic, Mehmed., Yin, Qi., & Liu, Xin. (2021). A Weight Possibilistic Fuzzy C-Means Clustering Algoritm. Hindawi: Scientific Programming Vol. 2021, Article ID 9965813, 10

[19] R. Krishnapuram and J. Keller. (1993). A possibilistic approch to clustering. IEEE TFS, vol. 1, no. 2, pp. 88–110

[20] Nurjanah, Farmadi, Andi., & Indriani, Fatma. (2014). Implementasi Metode Fuzzy C-Means Pada Sistem Clustering data varietas padi. Kumpulan Jurnal Ilmu Komputer (KLIK) Vol. 01 No.01 ISSN: 2406-7857

[21] Kusumadewi, Sri., & Purnomo, Hari. (2010). Aplikasi Logika Fuzzy Untuk Pendukung Keputusan. Yogyakarta, Graha Ilmu

[22] Setyawan, Andy Arief., & Ilham, Ahmad. (2019). A Novel Framework of the *Fuzzy C-Means* Distances Problem Based Weighted Distance. Journal of Applied Computing and Informatics

[23] Mattjik A A and Sumertajaya I M. (2011). Sidik Peubah Ganda. (Bogor: IPB Press)

[24] Timm N H. (2000). Applied Multivariate Analysis. (New York: Springer)

[25] Jollife I T. (2002). Principal Component Analysis. (New York: Springer)

[26] H, Dafitri., MS, Asih., & RI, Astuti. (2019). Media interaktif pengenalan angka dengan jari tangan menggunakan metode PCA. Journal of Information System Vo. 3 No. 2

[27] Adiyanto, Anggoro Teguh., UN, Dewi Handayani. (2022). Information Retrieval Sistem Kearsipan Pencarian Dokumen Di Dinas Pemberdayaan Perempuan dan Perlindungan Anak Kota Semarang Menggunakan Metode *Vector Space Model*. Jurnal Mahajana Informasi Vol. 7 No. 1 e-ISSN: 2527-8290

[28] S, A Yadav., & A, Sohal. (2017). Review paper on big data analytics in Cloud computing. Int J Comp Trends Technol (IJCTT) IX. 49(3);156-160

[29] Arbain, A. (2022). Komparasi Implementasi Model Machine Learning Hoax News Pada Local Dan Cloud Computing Deployment Menggunakan Google App Engine. Jurnal Informatika dan Teknik Elektro Terapan, 10(3)

[30] Berisha, B., Mëziu, E., & Shabani, I. (2022). Big data analytics in Cloud computing: an overview. J Cloud Comp 11, 24. https://doi.org/10.1186/s13677-022-00301-w

[31] Ghahremani-Nahr, Javid., & Nozari, Hamed. (2021). A Survey for Investigating Key Performance Indicators in Digital Marketing. International Journal of Innovation in Marketing Elements, 1(1), 1–6