Pengelompokan Data Jenis Kejahatan di Indonesia Menggunakan Metode *Agglomerative Hierarchical Clustering* (AHC) pada Tahun 2021

Faustina Alifah Mardhiyah*, Marizsa Herlina

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*faustinalifah@gmail.com, Marizsa.herlina@unisba.ac.id

Abstract. The development of technology makes it easier for the public to receive information, one of which is news abour crimes tha are occurring. This causes public unrest because many lives have been lost so people feel unsafe. This research aims to make it easier for law enforcement and the public to anticipate criminal acts. To find out the results of these objectives, the *Agglomerative Hierarchical Clustering* (AHC) method is used, where grouping is caried out to determine crime-prone areas. The data source used is the result of publications issued by the Central Statistics Agency. The data used is data on the number of crimes in 2021 consisting of 34 regional police forces in Indonesia where the data used is carried out by simple imputation first using the average due to missing data. Based on the results of cluster analysis of single linkage, complete linkage, average linkage, ward's method and centroid method, 2 clusters were obtained each. The results of the cluster analysis were compared eith the k-means method which esulted in the AHC method being the best mehod eith better index values compared to k-means.

Keywords: Agglomerative Hierarchical Clustering, Missing Data, Types of Crime, K-Means, Clustering.

Abstrak. Berkembangnya teknologi memudahkan masyarakat dalam menerima informasi, salah satunya adalah pemberitaan mengenai tindakan kejahatan yang sedang terjadi. Hal tersebut membuat keresahan masyarakat dikarenakan banyaknya nyawa yang hilang sehingga masyarakat merasa tidak aman. Tujuan dari penelitian ini adalah untuk memudahkan penegak hukum dan masyarakat dalam melakukan antisipasi tindakan kejahatan. Untuk mengetahui hasil dari tujuan tersebut digunakanlah metode *Agglomerative Hierarchical Clustering* (AHC) dimana dilakukan pengelompokan dalam menentukan daerah rawan kejahatan. Sumber data yang digunakan merupakan hasil publikasi yang dikeluarkan oleh Badan Pusat Statistik. Data yang digunakan adalah data jumlah kejahatan tahun 2021 ynag terdiri dari 34 kepolisian daerah di Indonesia dimana data yang digunakan dilakukan imputasi sederhana terlebih dahulu menggunakan rata-rata dikarenakan adanya data *missing*. Berdasarkan hasil analisis *cluster single linkage*, *complete linkage*, *average linkage*, *ward's method* dan *centroid method* masing-masing diperoleh 2 *cluster*. Hasil analisis *cluster* tersebut dibandingkan dengan metode *k-means* yang menghasilkan bahwa metode AHC merupakan metode terbaik dengan nilai indeks yang lebih baik dibandingkan dengan *k-means*.

Kata Kunci: Agglomerative Hierarchical Clustering, Data Missing, Jenis Kejahatan, K-Means, Pengelompokan.

Corresponding Author
Email: Marizsa.herlina@unisba.ac.id

A. Pendahuluan

Pada akhir tahun 2020, sekitar 1% populasi dunia (82,4 juta orang) telah dipindahkan secara paksa karena penganiayaan, konflik, atau kekerasan umum. Pandemic COVID-19 telah mengungkap dan memperkuat ketidaksetaraan dan diskriminasi. Nyatanya, krisis sangat mengganggu fungsi pemerintah, menguji, merusak, dan terkadang menghancurkan system hak dan perlindungan negara. Pemulihan krisis dan pembangunan berkelanjutan harus dibangun diatas perdamaian, stabilitas dan penghirmatan terhadap hal asasi manusia [1].

Kejahatan adalah perbuatan yang melanggar hukum, agama, atau norma sosial dan merupakan perbuatan yang merugikan atau menyimpang dari orang lain [2]. Menurut Badan Pusat Statistik (BPS), tingkat kejahatan di Indonesia mengalami penurunan pada tahun 2021. Jumlah kejahatan yang terjadi di Indonesia pada tahun 2021 sebanyak 239.481 kasus. Jumlah ini turun 3,135 dibanding tahun sebelumnya sebanyak 247.218 kasus [3]. Tindak kejahatan di Indonesia merupakan masalah umum diberbagai daerah. Hal ini sulit untuk memnentukan provinsi mana yang memiliki tingkat kerentanan tertentu terhadap kejahatan. Oleh karena itu, untuk mempermudah pemerintah dan kepolisian Republik Indonesia dalam menangani kejahatan maka diperlukan peng*cluster*an/pengelompokan suatu wilayah.

Dalam kasus ini, metode *cluster* yang digunakan adalah *Agglomerative Hierarchical Clustering* (AHC) yang terdiri dari *single linkage*, *complete linkage*, *average linkage*, *ward's method*, dan *centroid method*. Metode AHC memiliki suatu pendekatan salah satunya adalah *bottom-up* yaitu proses pengelompokan data yang dimulai dari pengelompokan terkecil sampai terbesar. Selanjutnya, tujuan dalam penelitian ini diuraikan dalam pokok-pokok sebagai berikut:

1. Untuk membuat pengelompokan terbaik dari metode *Agglomerative Hierarchical Clustering* (AHC).
2. Untuk mengetahui daerah yang termasuk rawan kejahatan di Indonesia berdasarkan hasil metode terbaik.

B. Metodologi Penelitian

Penelitian ini menggunakan data bersifat sekunder yang bersumber dari publikasi Statistika Kriminal 2022 yang dikeluarkan oleh Badan Pusat Statistik (BPS). Dengan variabel penelitian adalah data jumlah kejahatan berdasarkan jenis kejahatan di Indonesia tahun 2021 di 34 kepolisian daerah di Indonesia. Analisis data dilakukan dengan menggunakan *software* Microsoft Excel dan R Studio.

Langkah Analisis Data

Sebelum memulai tahapan analisis *cluster*, dilakukan penanganan data *missing* untuk mengisi data yang kosong dengan memprediksi nilai sehingga dapat dilakukan analisis yang baik.

1. Melakukan analisis cluster hierarchy menggunakan Agglomerative Hierarchical Clustering (AHC)
   Mengukur kemiripan antar objek menggunakan jarak *Euclidean* [4].
   $$d_{(xy)} = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$$
2. Menggabungkan objek x dan y yang memiliki jarak terdekat kemudian digabungkan menjadi satu *cluster*. Dari klaster xy yang sudah terbentuk, hitung jarak dengan objek lainnya yang belum bergabung dengan menggunakan rumus sebagai berikut [5]:
   a. Single Linkage
   $$d_{(XY)Z} = \min\{d_{XZ}, d_{YZ}\}$$
   b. Complete Linkage
   $$d_{(XY)Z} = \max\{d_{XZ}, d_{YZ}\}$$
   c. Average Linkage
   $$d_{(XY)Z} = avr\{d_{XZ}, d_{YZ}\}$$
   d. Ward's Method
   $$d_{(XY)Z} = SSE\{d_{XZ}, d_{YZ}\} = \frac{[(n_X + n_Z)d_{XZ} + (n_Y + n_Z)d_{YZ}] - n_Z d_{XY}}{n_X + n_Y + n_Z}$$

e. Centroid Method

$$d_{(XY)Z} = \frac{n_X}{n_X + n_Y} d_{XZ} + \frac{n_Y}{n_X + n_Y} d_{YZ} - \frac{n_X n_Y}{(n_X + n_Y)^2}$$

Memperbaharui matriks ukuran jarak setiap mendapatkan penggabungan objek baru. Melakukan tahapan secara berulang hinga semua objek tidak dapat berpindah lagi dan bergabung menjadi satu *cluster* besar.

3. Melakukan analisis *cluster non-hierarchy* menggunakan *k-means* dimana k *cluster* yang digunakan hasil dari *Silhouette* plot. Menentukan nilai *centroid* awal secara *random*. Kemudian, menghitung jarak setiap objek ke *centroid* dengan menggunakan jarak Euclidean dan dikelompokan berdasarkan jarak terdekat. Menghitung nilai *cenroid* baru dengan menggunakan rumus [6]:

$$v = \frac{\sum_{i=1}^{n} x_i}{n}; i = 1,2,3, \dots, n$$

Mengulangi langkah sampai *centroid* dan anggota *cluster* tidak dapat berpindah lagi.

4. Memilih metode terbaik dengan menggunakan rumus simpangan baku [7]:

$$\sigma = \frac{\sigma_w}{\sigma_b} \times 100\% = \frac{\frac{1}{K}\sum_{k=1}^{k} \sigma_k}{\left[\frac{1}{K}\sum(\mu_k - \mu)^2\right]^{\frac{1}{2}}} \times 100\% \text{ dimana, } \sigma_k = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_i - \mu_k)^2}$$

5. Membuat pemetaan dari metode terbaik dengan menggunakan *heat map* untuk mengetahui kepolisian daerah yang termasuk daerah rawan kejahatan di Indonesia.

C. Hasil Penelitian dan Pembahasan

Penanganan Data *Missing*

Melakukan penanganan data *missing* dalam suatu variabel dengan rata-rata dari semua nilai yang diketahui pada suatu variabel sebagai berikut [8]:

1. Imputasi data pada variabel kejahatan terhadap nyawa

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{29+96+10+\cdots+31}{33} = 28$$

2. Imputasi data pada variabel kejahatan terhadap kemerdekaan orang kejahatan

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{23+86+48+\cdots+148}{32} = 52$$

3. Imputasi data pada variabel kejahatan terhadap ketertiban umum

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{6 + 23 + 1 + \cdots + 348}{25} = 101$$

Mengukur Jarak Antar Objek

Untuk mengukur kesamaan antar objek menggunakan jarak Euclidean untuk mencari selisih antar data. Misalnya akan dihitung jarak dari data pertama (Aceh) dengan data kedua (Sumatera Utara) sebagai berikut [4]:

$$d_{1,2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_9 - y_9)^2}$$
$$d_{1,2} = \sqrt{(29 - 96)^2 + (1074 - 4287)^2 + \cdots + (6 - 23)^2}$$
$$d_{1,2} = \sqrt{153672624}$$
$$d_{1,2} = 12396{,}47627$$

Perhitungan jarak antara Aceh dengan Sumatera Utara menghasilkan jarak Euclidean sebesar 12396,47627 dan seterusnya hingga data terakhir yaitu data ke-34. Berikut ringkasan matriks dari hasil perhitungan jarak Euclidean:

$$D = \begin{bmatrix} 0 & 12396{,}476 & 603{,}279 & 648{,}882 & \dots & 1442{,}036 \\ 12396{,}476 & 0 & 12850{,}838 & 12144{,}831 & \dots & 12634{,}191 \\ 603{,}279 & 12850{,}838 & 0 & 1147{,}611 & \cdots & 1546{,}262 \\ 648{,}882 & 12144{,}831 & 1147{,}611 & 0 & \cdots & 1640{,}757 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1442{,}036 & 12634{,}191 & 1546{,}262 & 1640{,}757 & \cdots & 0 \end{bmatrix}$$

Hasil Pengelompokan Metode Hierarki

Setelah dilakukan perhitungan jarak, selanjutnya adalah menemukan data terdekat dalam matriks jarak Euclidean dan menggabungkannya menjadi *cluster* baru. Berdasarkan hasil matriks jarak, didapatkan jarak terdekat pada data ke-17 (Bali) dengan data ke-21 (Kalimantan Tengah).

$$\boldsymbol{Min(D_{Euc}) = \min(d_{17,21}) = 141,234}$$

Didapatkan jarak antara Bali dengan Kalimantan Tengah sebesar 141,234 yang kemudian kedua data tersebut digabungkan menjadi satu kelompok. Hal ini merupakan tahap awal untuk setiap metode yang akan digunakan pada penelitian ini.

Hasil *Cluster Metode Single Linkage, Complete Linkage, Average Linkage,* dan *Centroid Method*

Sebelumnya, pada tahap awal sudah didapatkan jarak terdekat yaitu data ke-17 (Bali) dengan data ke-21 (Kalimantan Tengah) sebesar 141,234. Selanjutnya adalah menghitung jarak antar kelompok (17 dan 21) dengan kelompok lain yang tersisa yaitu kelompok 1,2,3 dan seterusnya. Sehingga, didapatkan hasil pengelompokan kepolisian daerah di Indonesia berdasarkan jumlah kejahatan dengan memperolah 2 *cluster* sebagai berikut:

Tabel 1. Hasil Pengelompokan Single Linkage, Complete Linkage, Average Linkage, dan Centroid Method

Cluster	**Kepolisian Daerah**	**Jumlah**
1	Metro Jaya, Jawa Timur, Sulawesi Selatan, Sumatera Selatan, Papua, Jawa Barat, Lampung, Jawa Tengah, Riau, Aceh, Sumatera Barat, Sulawesi Tengah, Nusa Tenggara Timur, Sulawesi Utara, Kalimantan Barat, Sulawesi Tenggara, Gorontalo, Jambi, Nusa Tenggara Barat, Papua Barat, Banten, DI Yogyakarta, Bengkulu, Kepulauan Riau, Bali, Kalimantan Tengah, Maluku, Maluku Utara, Sulawesi Barat, Kep. Bangka Belitung, Kalimantan Utara, Kalimantan Selatan, dan Kalimantan Timur	33
2	Sumatera Utara	1

Hasil *Cluster Ward's Method*

Hasil *cluster* kepolisian daerah di Indonesia berdasarkan jumlah kejahatan dengan menggunakan *ward's method* memperoleh 2 *cluster* yang dapat dilihat sebagai berikut:

Tabel 2. Hasil Pengelompokan *Ward's Method*

Cluster	**Kepolisian Daerah**	**Jumlah**
1	Kalimantan Selatan, Kalimantan Timur, Maluku Utara, Sulawesi Barat, Kep. Bangka Belitung, Kalimantan Utara, Sulawesi Tenggara, Gorontalo, Maluku, Papua Barat, Nusa Tenggara Timur, Sulawesi Utara, Kalimantan Barat, Jambi, Nusa Tenggara Barat, Bali, Kalimantan Tengah, Bengkulu, Kepulauan Riau, DI Yogyakarta, Banten, Sumatera Selatan, Lampung, Jawa Tengah, Aceh, Riau, Papua, Jawa Barat, Sumatera Barat, dan Sulawesi Tengah	30
2	Sumatera Utara, Metro Jaya, Jawa Timur, dan Sulawesi Selatan	4

Hasil *Cluster K-Means*

Hasil *cluster* kepolisian daerah di Indonesia berdasarkan jumlah kejahatan dengan menggunakan *k-means* memperoleh 2 *cluster* yang dapat dilihat sebagai berikut:

Tabel 3. Hasil Pengelompokan *K-Means*

Cluster	**Kepolisian Daerah**	**Jumlah**
1	Aceh, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep. Bangka Belitung, Kepulauan Riau, Jawa Barat, Jawa Tengah, DI Yogyakarta, Banten, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, dan Papua	31
2	Sumatera Utara, Metro Jaya, dan Jawa Timur	3

Pemilihan Metode Terbaik

Simpangan Baku Metode *Single Linkage, Complete Linkage, Average Linkage*, dan *Centroid Method*

Diperoleh hasil simpangan baku sebagai berikut:

$$\boldsymbol{\sigma = \frac{\sigma_w}{\sigma_b} \times 100\% = \frac{218{,}859}{1387{,}210} \times 100\% = 15{,}78\%}$$

Simpangan Baku *Ward's Method*

Diperoleh hasil simpangan baku sebagai berikut:

$$\boldsymbol{\sigma = \frac{\sigma_w}{\sigma_b} \times 100\% = \frac{511{,}464}{819{,}644} \times 100\% = 62{,}04\%}$$

Simpangan Baku Metode *K-Means*

Diperoleh hasil simpangan baku sebagai berikut:

$$\sigma = \frac{\sigma_w}{\sigma_b} \times 100\% = \frac{489{,}017}{945{,}991} \times 100\% = 51{,}69\%$$

Berdasarkan hasil perhitungan nilai simpangan baku tersebut, dapat dilihat pada tabel sebagai berikut:

Tabel 4. Perbandingan Nilai Rasio Simpangan Baku

Metode	Nilai Rasio Simpangan Baku
Single Linkage	15,78%
Complete Linkage	15,78%
Average Linkage	15,78%
Ward's Method	62,40%
Centroid Method	15,78%
K-Means	51,69%

Berdasarkan Tabel 5 metode dengan nilai rasio simpangan baku terkecil, sehingga metode terbaik atau yang paling optimal adalah metode *Single Linkage*, *Complete Linkage*, *Average Linkage*, dan *Centroid Method* dengan nilai rasio simpangan baku sebesar 15,78%.

Pemetaan Hasil Metode Terbaik

Tabel 6. Nilai Persentase Metode Single Linkage, Complete Linkage, Average Linkage, dan Centroid Method pada Masing-Masing Variabel Disetiap Cluster

Variabel	*Cluster* 1	*Cluster* 2
Kejahatan Terhadap Nyawa	21%	79%
Kejahatan Terhadap Fisik	14%	86%
Kejahatan Terhadap Kesusilaan	14%	86%
Kejahatan Terhadap Kemerdekaan Orang Kejahatan	37%	63%
Kejahatan Terhadap Hak Milik/Barang dengan Penggunaan Kekerasan	19%	81%
Kejahatan Terhadap Hak Milik/Barang	12%	88%
Kejahatan Terkait Narkotika	14%	86%
Kejahatan Terkait Penipuan, Penggelapan dan Korupsi	14%	86%
Kejahatan Terhadap Ketertiban Umum	82%	18%

Berdasarkan Tabel 6 dapat disimpulkan dimana kepolisian daerah pada *cluster* 1 menunjukkan tingkat kejahatan pada daerah tersebut rendah atau dikatakan kepolisian daerah pada *cluster* 1 aman dengan dilihat dari nilai *cluster* pada setiap variabelnya yang rendah. Walaupun pada *cluster* 1 termasuk kedalam kategori yang rendah, perlu di waspadai untuk jenis kejahatan terhadap ketertiban umum karena di wilayah yang termasuk dalam *cluster* 1 angka kejahatan tersebut merupakan kejahatan paling tinggi sebesar 82%. Sedangkan untuk kepolisian daerah pada *cluster* 2 dilihat dari nilai *cluster* pada setiap variabelnya menunjukkan tingkat kejahatan yang tinggi dibandingkan *cluster* 1 atau dapat dikatakan kepolisian daerah pada *cluster* 2 dikategorikan rawan kejahatan. Untuk jenis kejahatan tertinggi pada *cluster* 2 yaitu kejahatan terhadap hak milik/barang sebesar 88% sehingga perlu dilakukan pengawasan ekstra.

Kelompok yang terbentuk divisualisasikan dalam bentuk *heat map* seperti pada Tabel 6 terdapat dua daerah rawan kejahatan, yaitu aman (rendah) pada *cluster* 1 dan rawan (tinggi) pada *cluster* 2. Wilayah dengan tingkat kejahatan yang rendah memiliki warna krem sedangkan wilayah dengan tingkat kejahatan yang tinggi memiliki warna merah. Dapat dikatakan bahwa

semakin terang warnanya maka semakin aman wilayah tersebut dan sebaliknya.

D. Kesimpulan

Berdasarkan pembahasan dalam penelitian ini, peneliti menyimpulkan beberapa hasil penelitian sebagai berikut:

1. Berdasarkan hasil perbandingan *cluster* menggunakan *Agglomerative Hierarchical Clustering* (AHC), didapatkan metode *single linkage*, *complete linkage*, *average linkage*, dan *centroid method* menjadi pengelompokan terbaik pada data jumlah kejahatan di Indonesia tahun 2021.
2. Pengelompokan terbaik yang didapatkan adalah metode AHC dimana terdapat satu perbedaan dalam pemetaan daerah rawan kejahatan. Pada *single linkage*, *complete linkage*, *average linkage*, dan *centroid method* kepolisian daerah Metro Jaya, Jawa Timur, dan Sulawesi Selatan masuk ke dalam *cluster* 1 dan sebaliknya untuk *ward's method*.

Acknowledge

Peneliti mengucapkan terima kasih kepada semua pihak yang sudah membantu menyelesaikan dan memberi saran untuk penelitian ini.

Daftar Pustaka

[1] Kementrian PPN/Bappenas. Sustainable Development Goals. https://sdgs.bappenas.go.id/.

[2] Dewi, S. M., Windarto, A. P., Damanik, I. S., & Satria, H. (2019). Analisa Metode K-Means pada Pengelompokan Kriminalitas Menurut Wilayah. *Seminar Nasional Sains &Teknologi Informasi (SENSASI)*, 620-625.

[3] BPS. (2022). *Statistik Kriminal 2022*. Jakarta: Badan Pusat Statistik.

[4] Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. *Fifth Edition*. New Jersey: Prentice Hall Inc.

[5] Hair, J. F. Jr., R. E. Anderson, R. L. Thatham, & W. C. Black. (2010). *Multivariate Data Analysis Fifth Edition*. New Jersey: Prentice Hall International, Inc.

[6] Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: Penerbit Andi.

[7] Bunkers, W. J., Miller, J. R., & DeGaetano, A. T. (1996). *Definition of Climate Regions in the Northern Plains Using an Objective Cluster Modification Technique*.

[8] Acuna, E. & Rodrigues, C. (2004). *The Treatment of Missing Values and its Effect is the Classifier Accuracy*. *Proceedings of the Meeting of the International Federation of Classification Sociaties (IFCS)*.

[9] Agnesya Risnandar, & Anneke Iswani Achmad. (2023). Pemodelan Generalized Space Time Autoregressive untuk Meramalkan Indeks Harga Konsumen. *Jurnal Riset Statistika*, 43–50. https://doi.org/10.29313/jrs.v3i1.1792

[10] Nur, F., 1∗, A., & Achmad, A. I. (2023). Perbandingan Fuzzy C-Means Clustering dan Fuzzy Possibilistic C-Means Clustering dalam Pengelompokan Kabupaten/Kota di Jawa Barat Berdasarkan Akses terhadap Sumber Air dan Sanitasi Layak Pada Tahun 2020. 1(1), 27–34. https://doi.org/10.29313/datamath.v1i1.16

[11] Salsabila Pratiwi, & Marizsa Herlina. (2023). Pengaruh Harga Pangan terhadap Inflasi dengan Metode Vector Autoregressive Integrated Moving Average. *Jurnal Riset Statistika*, 87–96. https://doi.org/10.29313/jrs.v3i2.2690