

# Klasterisasi Provinsi di Indonesia Berdasarkan Angka Harapan Hidup Menggunakan *K-Means* dengan Evaluasi *Elbow Method*

Asep Lukman Arip Hidayat<sup>1</sup>, Helmi Zulqan<sup>2</sup>, Gandung Triyono<sup>3</sup>

<sup>1,2</sup> Magister Ilmu Komputer Fakultas Teknologi Informasi Universitas Budi Luhur Jakarta  
Jl. Ciledug Raya, Petukangan Utara, Kec. Pesangrahan, Kota Jakarta Selatan, DKI Jakarta

<sup>1</sup>aseplukman.ah@gmail.com

<sup>2</sup>hzulqan@gmail.com

<sup>3</sup>gandung.triyono@budiluhur.ac.id

## Abstract

Angka Harapan Hidup (AHH) adalah indikator penting untuk mengukur kesejahteraan masyarakat dan tingkat kesehatan di suatu daerah. AHH Indonesia sangat berbeda antar provinsi, yang menunjukkan perbedaan dalam infrastruktur dan akses layanan kesehatan. Tujuan dari penelitian ini adalah untuk menggunakan algoritma *K-Means* untuk meklasterisasi provinsi-provinsi di Indonesia berdasarkan AHH, dan untuk mengevaluasi hasil klasterisasi dengan metode *Elbow* untuk menentukan jumlah klaster yang ideal. Data yang digunakan mencakup AHH dari seluruh provinsi Indonesia dalam beberapa tahun terakhir. Untuk mempermudah identifikasi daerah yang memerlukan perhatian khusus dalam perencanaan kebijakan kesehatan, provinsi diklasifikasikan ke dalam klaster berdasarkan karakteristik AHH yang sebanding. Studi ini fokus pada optimasi Algoritma *K-Means* dengan metode *Elbow*. Percobaan iterasi dilakukan sepuluh kali dan menemukan nilai K ideal, yaitu K=3. Hasil klasterisasi menunjukkan bahwa Cluster 0 memiliki 103 anggota, Cluster 1 memiliki 181 anggota, dan Cluster 2 memiliki 260 anggota. Hasil penelitian menunjukkan bahwa algoritma *K-Means* berhasil mengelompokkan provinsi-provinsi berdasarkan AHH dengan tingkat variasi yang rendah di setiap klaster. Metode *Elbow* membantu menentukan jumlah klaster yang ideal.

**Keywords:** *Angka Harapan Hidup, Data Mining, Elbow Method, K-Means, Klasterisasi.*

## I. PENDAHULUAN

Angka Harapan Hidup (AHH) merupakan indikator utama yang digunakan untuk menilai tingkat kesejahteraan dan kesehatan masyarakat di suatu wilayah. AHH menjadi ukuran keberhasilan pemerintah dalam menyediakan pelayanan kesehatan yang memadai dan meningkatkan kualitas hidup masyarakat. Di Indonesia, rata-rata angka harapan hidup mencapai 73,5 tahun pada tahun 2021, yang menunjukkan adanya perbaikan dalam kesejahteraan masyarakat secara umum [1], [2]. Namun, terdapat ketimpangan AHH antar provinsi yang cukup besar, terutama karena perbedaan akses terhadap layanan dan fasilitas umum.

Penelitian terkait klasterisasi AHH sebelumnya telah dilakukan dengan menggunakan berbagai pendekatan, seperti algoritma *backpropagation* dan *Conjugate Gradient* [3], [4]. Penelitian-penelitian tersebut berhasil memetakan pola distribusi AHH di wilayah tertentu dan memberikan wawasan tentang pengaruh faktor sosial dan ekonomi. Namun, belum banyak penelitian yang secara spesifik menggunakan algoritma *K-Means* untuk menganalisis persebaran AHH secara nasional dengan evaluasi menggunakan metode *Elbow*. Pendekatan ini penting untuk memberikan gambaran lebih jelas mengenai kelompok-kelompok provinsi berdasarkan karakteristik AHH-nya [1], [2]. Distribusi AHH antar provinsi masih menunjukkan kesenjangan yang signifikan akibat perbedaan akses terhadap

fasilitas kesehatan dan infrastruktur pendukung lainnya [2]. Penelitian ini bertujuan untuk melakukan klasterisasi provinsi-provinsi di Indonesia berdasarkan angka harapan hidup menggunakan algoritma *K-Means*, dengan metode *Elbow* digunakan untuk menentukan jumlah klaster yang optimal. Melalui analisis ini, diharapkan dapat diidentifikasi provinsi-provinsi dengan AHH rendah yang membutuhkan intervensi kebijakan lebih lanjut. Dengan cara ini, hasil penelitian dapat menjadi landasan dalam perencanaan pembangunan kesehatan yang lebih efisien dan terarah [2], [5]. Dalam penelitian ini, metode *Elbow* akan digunakan untuk mengevaluasi hasil klasterisasi. Diharapkan metode ini dapat membantu pemerintah membuat keputusan tentang pembangunan sosial dan intervensi kesehatan di daerah dengan angka harapan hidup yang rendah [2], [6].

Angka Harapan Hidup (AHH) adalah salah satu indikator penting untuk mengukur tingkat kesejahteraan dan kesehatan masyarakat. AHH diukur berdasarkan tahun rata-rata yang diharapkan akan dijalani oleh seorang bayi yang baru lahir dengan mempertimbangkan kondisi kesehatan saat itu [1], [2]. AHH menjadi tolok ukur penting dalam mengevaluasi kualitas pelayanan kesehatan, seperti akses ke fasilitas kesehatan, penanganan penyakit, dan upaya untuk meningkatkan gizi dan sanitasi [6]. Selain itu, AHH juga berfungsi sebagai salah satu faktor penting dalam mengevaluasi kualitas.

Studi sebelumnya telah menunjukkan bahwa AHH sangat penting untuk menentukan kebijakan kesehatan dan kesejahteraan. Menurut penelitian [1], prediksi AHH yang akurat dapat membantu pemerintah membuat keputusan berbasis data mengenai program kesejahteraan masyarakat nasional dan daerah. AHH di berbagai provinsi Indonesia sangat berbeda, yang dipengaruhi oleh perbedaan infrastruktur kesehatan, ekonomi, dan kondisi sosial [3], [4]. Oleh karena itu, pemetaan AHH menjadi langkah penting dalam mengidentifikasi wilayah yang membutuhkan perhatian lebih besar guna meningkatkan kesejahteraan masyarakat. Dalam penelitian ini, data yang digunakan mencakup periode 2016–2023, sehingga analisis yang dilakukan dapat mencerminkan tren terbaru dalam perubahan AHH di berbagai provinsi.

Salah satu metode yang paling umum digunakan dalam klusterisasi adalah metode jari-jari. Metode ini bekerja dengan menilai variasi dalam setiap kluster dan memplotnya terhadap jumlah kluster yang berbeda, sehingga titik optimal atau "jari-jari" pada grafik menunjukkan jumlah kluster yang menghasilkan variasi paling rendah tanpa *overfitting* [2], [6]. Dengan cara ini, peneliti dapat menentukan jumlah kluster yang optimal.

Menurut penelitian yang dilakukan oleh [6], metode jari-jari bersama dengan *K-Means* dapat meningkatkan efisiensi pemrosesan data yang kompleks seperti angka harapan hidup. Selain itu, metode jari-jari membantu menghindari pemilihan jumlah kluster yang terlalu sedikit atau terlalu banyak, yang dapat menyebabkan hasil analisis yang tidak representatif atau terlalu rumit [4].

Dengan menggunakan metode *K-Means*, jaringan syaraf tiruan, dan algoritma lainnya, sejumlah penelitian telah dilakukan di berbagai wilayah di Indonesia tentang klusterisasi angka harapan hidup. Misalnya, penelitian yang dilakukan oleh [3] mencoba memprediksi angka harapan hidup di Kota Jambi dengan menggunakan algoritma *backpropagation*. Hasil penelitian ini menunjukkan bahwa penggunaan metode pembelajaran mesin dapat membantu dalam perencanaan kebijakan kesehatan dan membuat prediksi yang akurat. Studi tambahan yang dilakukan oleh [4] di Jawa Timur menggunakan algoritma *Conjugate Gradient-Backpropagation*, yang menunjukkan hasil efektif dalam memprediksi angka harapan hidup. Hal ini membuktikan bahwa penggunaan metode *Machine Learning* dapat digunakan untuk mengolah data Angka Harapan Hidup (AHH).

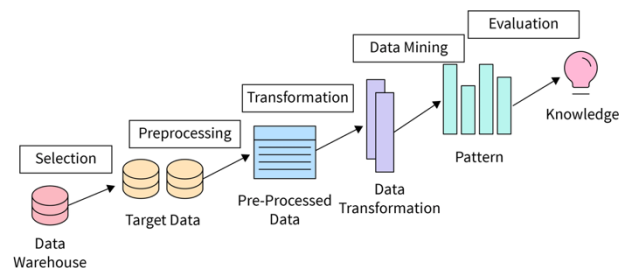
Penelitian ini bertujuan untuk mengelompokkan provinsi berdasarkan angka harapan hidup (AHH), terutama dengan memanfaatkan kombinasi algoritma *K-Means* dan metode evaluasi *Elbow*. Kontribusi dari penelitian ini adalah penggunaan data BPS Angka Harapan Hidup dengan rentang tahun 2016-2023 per provinsi di Indonesia. Diharapkan, hasil klusterisasi ini dapat memberikan pemahaman yang lebih jelas tentang pola distribusi AHH di Indonesia, sehingga dapat menjadi dasar bagi pembuat kebijakan dalam merancang strategi kesehatan yang lebih tepat sasaran. Dengan demikian, penelitian ini berkontribusi dalam mendorong penggunaan data sebagai dasar dalam perencanaan kebijakan kesehatan yang lebih efektif dan terarah.

## II. METODOLOGI PENELITIAN

Bagian ini menjelaskan secara rinci tentang penelitian yang dilakukan mulai dari langkah-langkah sampai metode yang digunakan.

### A. Knowledge Discovery in Database (KDD)

Pada metodologi penelitian terdapat susunan kerangka kerja yang akan diikuti. Adapun kerangka kerja yang dilakukan pada penelitian ini adalah langkah-langkah *Knowledge Discovery in Database* atau biasa disebut dengan KDD, dapat dilihat pada Gambar 1.



Gambar 1. Flowchart Tahapan KDD [7]

Gambar 1. menunjukkan tahapan-tahapan dalam proses KDD, yaitu:

- 1) *Pemilihan Data (Data Selection)*: Pada tahap ini, data yang relevan dipilih agar dapat diolah atau digunakan dalam penelitian.
- 2) *Pra-pemrosesan Data (Data Preprocessing)*: Data yang telah dipilih dibersihkan dari gangguan, seperti noise, inkonsistensi, atau data yang hilang, agar kualitasnya lebih baik dan siap digunakan.
- 3) *Transformasi Data (Data Transformation)*: Data yang sudah bersih kemudian diubah ke dalam format yang sesuai untuk memudahkan analisis, misalnya melalui normalisasi atau pembuatan fitur baru berdasarkan data yang ada.
- 4) *Penambangan Data (Data Mining)*: Tahap ini melibatkan penggunaan teknik atau algoritma tertentu untuk menemukan pola atau informasi bermakna dari data yang telah diproses sebelumnya.
- 5) *Interpretasi dan Evaluasi (Interpretation and Evaluation)*: Hasil model dievaluasi untuk memastikan relevansi dan validitasnya. Informasi atau pola yang ditemukan kemudian diinterpretasikan agar dapat diterapkan dalam kehidupan nyata.

### B. Klusterisasi dengan Algoritma K-Means

*K-Means*, algoritma klusterisasi non-hierarki, mengelompokkan data ke dalam berbagai kluster berdasarkan kemiripan karakteristik. Algoritma ini bekerja dengan membagi data menjadi sejumlah kluster yang ditentukan yang didasarkan pada *centroid* atau titik tengah setiap kluster. Kluster-kluster ini kemudian diperbarui secara *iteratif* hingga variasi data dalam satu kluster seminimal mungkin [1], [4]. Penggunaan algoritma *K-Means* dalam pengelompokan angka harapan hidup memungkinkan untuk mengidentifikasi provinsi-provinsi di Indonesia yang memiliki kesamaan dalam. Klusterisasi ini membantu pemerintah membuat kebijakan kesehatan yang lebih terfokus dan berhasil.

Dalam beberapa penelitian, algoritma *K-Means* telah digunakan untuk berbagai analisis, salah satunya adalah pengelompokan provinsi berdasarkan data sosial-ekonomi dan demografi. Penelitian oleh [2] menggunakan algoritma ini untuk mengelompokkan provinsi di Indonesia berdasarkan angka harapan hidup dari tahun 2019 hingga 2021, dan berhasil menemukan pola persebaran yang menunjukkan provinsi dengan angka harapan hidup tertinggi dan terendah. Penelitian oleh [5] tentang penerapan algoritma *K-Means Clustering* untuk memetakan data desa berdasarkan tingkat penyebaran penyakit DBD. Data yang digunakan pada penelitian tersebut adalah data penyebaran DBD di Kecamatan Setia Janji. Hasil penelitian tersebut menunjukkan kluster sebanyak 3 kluster yang diukur dengan Davies-Bouldin Index (DBI). Penelitian oleh [8] yang berjudul “Penerapan Algoritma *K-Means* dalam Mengelompokkan Balita yang Mengalami Gizi Buruk Menurut Provinsi”. Penelitian tersebut bertujuan untuk mengelompokkan provinsi berdasarkan tingkat balita yang mengalami gizi buruk menggunakan algoritma *K-Means*.

Klasterisasi menggunakan algoritma *K-Means* dilakukan dalam langkah-langkah berikut [9] :

1) *Pemilihan Centroid Awal*: Pada tahap ini dilakukan pemilihan nilai  $k$  data dari  $n$  sampel sebagai centroid awal secara acak. Titik tersebut akan menjadi titik awal untuk pembentukan cluster.

2) *Menentukan Kluster*: Tahap ini menghitung jarak setiap data ke titik centroid yang sudah ditentukan. perhitungan tersebut dilakukan dengan Persamaan 1.1.

$$d(x_i, \mu_j) = \sqrt{\sum_{k=1}^n (x_{ik} - \mu_{jk})^2} \quad (1.1)$$

3) *Menghitung Centroid Baru*: Setelah kluster terbentuk, selanjutnya adalah menghitung centroid baru dengan rata-rata semua data yang termasuk di dalam kluster. Perhitungan tersebut dapat dilihat pada Persamaan 1.2.

$$\mu_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i \quad (1.2)$$

4) *Proses Iterasi*: Lakukan proses 2 dan 3 berulang sampai titik centroid tidak berubah (konvergen).

### III. HASIL DAN PEMBAHASAN

Hasil klasterisasi pada data angka harapan hidup bertujuan untuk mengelompokkan wilayah berdasarkan kesamaan tertentu, seperti tingkat harapan hidup tinggi, sedang, dan rendah. Analisis ini memberikan gambaran pola kelompok yang terbentuk. Pembahasan hasil ini diharapkan memberikan wawasan mendalam mengenai karakteristik tiap kelompok, membantu dalam merumuskan kebijakan yang tepat, serta mengidentifikasi langkah-langkah perbaikan bagi wilayah dengan angka harapan hidup lebih rendah.

#### A. Data Selection

*Dataset* yang digunakan pada penelitian ini adalah *Dataset* Angka Harapan Hidup berdasarkan Provinsi di Indonesia.

*Dataset* ini perlu dilakukan *preprocessing* agar data diolah dengan baik.

| Provinsi Angka Harapan Hidup (AHH) Menurut Provinsi dan Jenis Kelamin (Tahun) |                      |           |           |           |       |
|-------------------------------------------------------------------------------|----------------------|-----------|-----------|-----------|-------|
|                                                                               |                      | Unamed: 2 | Unamed: 3 | Unamed: 4 |       |
| 0                                                                             | NaN                  | NaN       | NaN       | NaN       |       |
| 1                                                                             | NaN                  | Laki-laki | NaN       | Perempuan | NaN   |
| 2                                                                             | NaN                  | 2022      | 2023      | 2022      | 2023  |
| 3                                                                             | ACEH                 | 68,26     | 68,42     | 72,16     | 72,35 |
| 4                                                                             | SUMATERA UTARA       | 67,66     | 68,06     | 71,60     | 72,00 |
| 5                                                                             | SUMATERA BARAT       | 67,99     | 68,27     | 71,89     | 72,21 |
| 6                                                                             | RIAU                 | 70,03     | 70,28     | 73,90     | 74,29 |
| 7                                                                             | JAMBI                | 69,57     | 69,83     | 73,49     | 73,81 |
| 8                                                                             | SUMATERA SELATAN     | 68,38     | 68,74     | 72,29     | 72,68 |
| 9                                                                             | BENGKULU             | 67,74     | 67,99     | 71,68     | 71,94 |
| 10                                                                            | LAMPUNG              | 69,07     | 69,31     | 72,97     | 73,28 |
| 11                                                                            | KEP. BANGKA BELITUNG | 69,06     | 69,30     | 72,97     | 73,26 |
| 12                                                                            | KEP. RIAU            | 68,58     | 68,99     | 72,47     | 72,94 |
| 13                                                                            | DKI JAKARTA          | 71,45     | 71,88     | 75,22     | 75,61 |
| 14                                                                            | JAWA BARAT           | 71,77     | 72,01     | 75,48     | 76,19 |

**Gambar 2.** Dataset Angka Harapan Hidup Provinsi di Indonesia

Pada Gambar 2. *Data Selection* dilakukan dengan memilih empat atribut yaitu Nama Provinsi, Jenis Kelamin, dan Angka Harapan Hidup dan Tahun.

#### B. Preprocessing Data

Data preprocessing mencakup metode seperti pembersihan data, transformasi data, dan reduksi data, yang semuanya berperan dalam meningkatkan kualitas data sebelum tahap analisis atau penambangan data [10]. Data akan diperiksa apakah memiliki nilai *missing* atau tidak. *Missing value* yang termasuk kedalam proses Data Cleaning seperti baris yang tidak perlu atau kosong tanpa data..

#### C. Data Transformation

Pada tahap transformasi data, *Dataset* yang ada akan diubah menjadi sebuah *Dataset* baru dengan format yang berbeda dari *Dataset* sebelumnya. Transformasi data merupakan tahap di mana data diubah ke dalam bentuk yang sesuai dengan kebutuhan analisis atau model yang akan digunakan [11]. Dalam penelitian ini, *Dataset* memerlukan transformasi, sehingga proses dapat dilanjutkan ke tahap berikutnya, yaitu *data mining*.

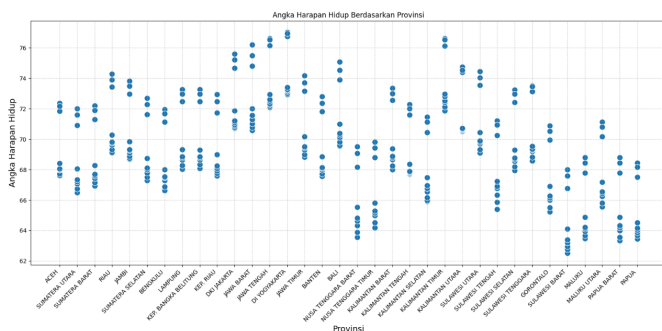
```
Index: 544 entries, 0 to 558
Data columns (total 4 columns):
#   Column                Non-Null Count  Dtype
---  ---                ---
0   Nama Provinsi          544 non-null    object
1   Jenis Kelamin          544 non-null    int64
2   Angka Harapan Hidup    544 non-null    float64
3   Tahun                  544 non-null    int64
dtypes: float64(1), int64(2), object(1)
```

**Gambar 3.** Tampilan Informasi Atribut *Dataset*

#### D. Data Mining

Data mining dapat dipandang sebagai langkah dalam proses *Knowledge Discovery in Databases* (KDD), yang mencakup pembersihan data, integrasi, pemilihan data, transformasi, penambangan data, penilaian pola, dan penyajian data [12]. Hal yang dilakukan pada tahap ini yaitu *Data Mining Clustering* dengan algoritma *K-Means* yang kemudian di optimasi dengan

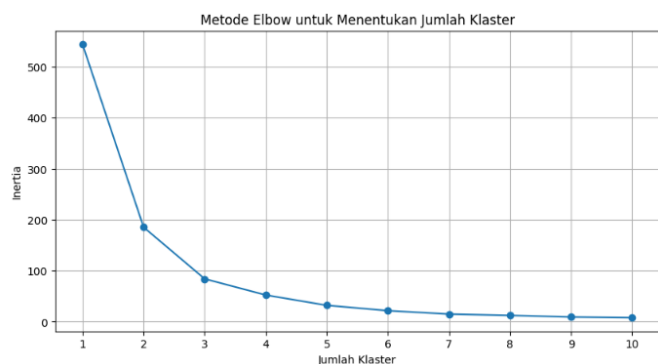
algoritma *Davies-Bouldin Index* untuk menentukan nilai K yang optimal.



**Gambar 4.** Tampilan Scatter Angka Harapan Hidup

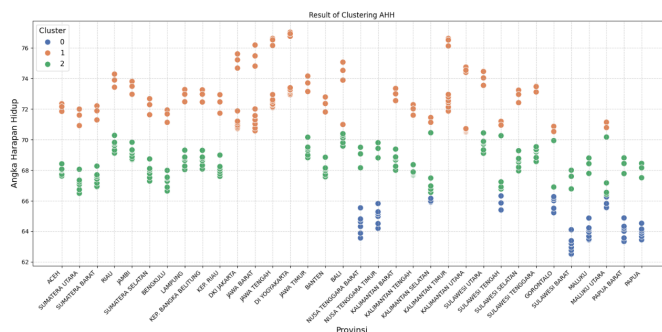
#### E. Evaluation

Tahap evaluasi ini melakukan pengecekan Nilai K Optimal Menggunakan *Elbow Method* yang dilakukan sebanyak 10 iterasi. Setelah iterasi selesai maka didapatkan nilai K yang optimal yaitu 3 cluster. Hasil iterasi dapat dilihat pada Gambar 5.



**Gambar 5.** Hasil Evaluasi *Elbow Method*

Setelah didapatkan nilai optimal, selanjutnya dilakukan proses *Fitting* dengan algoritma *K-Means* yang menghasilkan 3 Cluster dengan nilai *Davies-Bouldin Index* sebesar 0.47, *Silhouette Score* 0.59. Berikut ini tampilan *Scatter Plot* setelah dilakukan proses *clustering*.



**Gambar 6.** Tampilan Hasil *Clustering*

Gambar 6. menampilkan hasil klasterisasi yang menghasilkan tiga cluster yaitu Cluster 0, Cluster 1 dan Cluster 2. Cluster-cluster tersebut dapat dikategorikan dalam kategori “Rendah”,

“Menengah” dan “Tinggi”. Cluster 0 terdiri dari 103 anggota dengan nilai AHH dalam rentang nilai 62,49 sampai 66,32 sehingga Cluster 0 dikategorikan dalam cluster dengan nilai AHH “Rendah”. Cluster 1 terdiri dari 181 anggota dengan rentang nilai 70,49 sampai dengan 77,03. Cluster 1 dikategorikan dalam cluster dengan nilai AHH “Tinggi”. Cluster 2 yang terdiri dari 260 anggota yang memiliki nilai rentang AHH mulai dari 66,43 sampai dengan 70,45, sehingga cluster 1 dikategorikan pada cluster “Menengah”. Nilai Angka Harapan Hidup (AHH) terendah terdapat pada Provinsi Gorontalo pada tahun 2016, sedangkan untuk nilai Angka Harapan Hidup (AHH) tertinggi dimiliki oleh Provinsi Sumatera Utara pada tahun 2023.

#### IV. KESIMPULAN

Penelitian ini menitik beratkan pada optimasi Algoritma *K-Means* menggunakan *Elbow Method* untuk menentukan nilai K yang optimal. Percobaan iterasi dilakukan sebanyak 10 kali dan menghasilkan nilai K optimal yaitu K=3. Berdasarkan hasil penelitian ini dapat disimpulkan bahwa Cluster 0 terdiri dari 103 anggota dengan predikat rendah, Cluster 1 terdiri dari 181 anggota dengan predikat tinggi, dan Cluster 2 terdiri dari 260 anggota dengan predikat sedang. Beberapa saran penulis untuk penelitian selanjutnya agar dapat menggunakan algoritma *clustering* yang lain dan diharapkan hasil menjadi lebih baik karena ada perbandingan antara algoritma tersebut. Penelitian selanjutnya juga diharapkan menggunakan *dataset* yang lebih banyak dan variatif agar hasil yang diperoleh menjadi lebih baik.

#### V. SARAN

Penelitian ini memberikan kontribusi penting dalam mengelompokkan wilayah berdasarkan angka harapan hidup menggunakan metode klasterisasi, sehingga dapat mengungkap pola-pola yang terkait dengan kondisi sosial, ekonomi, dan kesehatan. Bagi peneliti selanjutnya, disarankan untuk mengeksplorasi faktor-faktor yang lebih spesifik, seperti akses terhadap layanan kesehatan, pendidikan, atau infrastruktur, yang mungkin memengaruhi hasil klasterisasi.

Selain itu, penggunaan metode klasterisasi yang berbeda atau penggabungan dengan analisis spasial dapat memberikan perspektif yang lebih komprehensif, sehingga mampu mendukung pengambilan keputusan yang lebih efektif dalam perencanaan kebijakan pembangunan wilayah.

#### REFERENSI

- [1] W. Kurniawan dan U. Indahyanti, “Prediksi Angka Harapan Hidup Penduduk Menggunakan Metode XGBoost,” *Indonesian Journal of Applied Technology*, vol. 1, no. 2, hlm. 18, Jul 2024, doi: 10.47134/ijat.v1i2.3045.
- [2] Y. Pratiwi dan Mulyawan, “Implementasi Algoritma K-Means untuk Menentukan Angka Harapan Hidup berdasarkan Tingkat Provinsi,” *Blend Sains Jurnal Teknik*, vol. 1, no. 4, hlm. 284–294, Mar 2023, doi: 10.56211/blendsains.v1i4.233.

- [3] P. Natalia Napitupulu, A. Rahim Damanik, J. Evonella Napitupulu, T. Informatika, S. A. Tunas Bangsa Jln Sudirman Blok No, dan S. Utara, "Implementasi Algoritma Backpropagation Jaringan Syaraf Tiruan untuk Prediksi Angka Harapan Hidup di Kota Jambi," Online, 2023.
- [4] S. Maria, S. Silalahi, dan L. Pujiastuti, "Penerapan Algoritma Conjugate Gradient Polak Ribiere Dalam Memprediksi Angka Harapan Hidup Di Jawa Timur," Online, 2023.
- [5] M. A. Sembiring, R. T. A. Agus, dan M. F. L. Sibuea, "Penerapan Metode Algoritma K-Means Clustering untuk Pemetaan Penyebaran Penyakit Demam Berdarah Dengue (DBD)," 2021. [Daring]. Tersedia pada: <http://jurnal.goretanpena.com/index.php/JSSR>
- [6] L. L. J. Anarki, Solikhun, dan F. Ilmi, "Optimisasi Prediksi Angka Harapan Hidup: Memanfaatkan Keunggulan Algoritma Conjugate Gradient Polak-Ribiere Di Sumatera Utara," Online, 2024.
- [7] Utkarsh, "Knowledge Discovery in Databases." Diakses: 4 Januari 2025. [Daring]. Tersedia pada: <https://www.scaler.com/topics/data-mining-tutorial/kdd-in-data-mining/>
- [8] M. D. Chandra, E. Irawan, I. S. Saragih, A. P. Windarto, dan D. Suhendro, "Penerapan Algoritma K-Means dalam Mengelompokkan Balita yang Mengalami Gizi Buruk Menurut Provinsi," *BIOS: Jurnal Teknologi Informasi dan Rekayasa Komputer*, vol. 2, no. 1, hlm. 30–38, Mar 2021, doi: 10.37148/bios.v2i1.19.
- [9] G. Feng, M. Fan, dan Y. Chen, "Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining," *IEEE Access*, vol. 10, hlm. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.
- [10] C. Li, "Preprocessing Methods and Pipelines of Data Mining: An Overview," Jun 2019, [Daring]. Tersedia pada: <http://arxiv.org/abs/1906.08510>
- [11] F. Alghifari dan D. Juardi, "Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," 2021.
- [12] D. N. Yoliadi, "Data mining Dalam Analisis Tingkat Penjualan Barang Elektronik Menggunakan Algoritma K-Means," 2023.