PENERAPAN ALGORITMA *K-MEANS* DALAM SEGMENTASI DAERAH BERDASARKAN JUMLAH WARGA YANG TERDAFTAR MENJADI PESERTA BPJS KESEHATAN DI JAWA BARAT

Zahratul Jannah, Agung Susilo Yuda Irawan, E.Haodudin Nurkifli
Program Studi Informatika, Fakultas Ilmu Komputer
Universitas Singaperbangsa Karawang, Indonesia
2010631170038@student.unsika.ac.id

ABSTRAK

Upaya meningkatkan akses terhadap layanan kesehatan yang terjangkau dan berkualitas merupakan agenda penting dalam mencapai *Universal Health Coverage* (UHC) secara global. Di Indonesia, Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan menjalankan program Jaminan Kesehatan Nasional (JKN) sebagai instrumen utama untuk mewujudkan UHC. Penelitian ini bertujuan untuk melakukan segmentasi wilayah di Jawa Barat berdasarkan jumlah penduduk yang terdaftar sebagai peserta BPJS Kesehatan pada tahun 2019-2023 dengan menggunakan algoritma *K-means* dan memvisualisasikan hasilnya. Penentuan jumlah klaster optimal dilakukan dengan metode *elbow* dan *silhouette coefficient*. Algoritma *K-means* menghasilkan tiga klaster wilayah, yaitu klaster 0 (501 kecamatan) dengan jumlah peserta BPJS tinggi, klaster 1 (222 kecamatan) dengan jumlah peserta BPJS menengah, dan klaster 2 (4 kecamatan) dengan jumlah peserta BPJS rendah. Evaluasi kinerja algoritma menunjukkan bahwa klaster terbaik adalah tiga klaster dengan nilai *Sum of square error* (SSE) terendah sebesar 1,67409 dan *Davies Bouldin Index* (DBI) terendah sebesar 0,36915605275672103. Visualisasi hasil segmentasi menggunakan *Quantum geographic information system* (QGIS) membantu mengidentifikasi wilayah prioritas yang memerlukan perhatian lebih dalam peningkatan pemanfaatan layanan BPJS. Penelitian ini memberikan informasi berharga bagi pemangku kepentingan, seperti pemerintah daerah dan BPJS Kesehatan, untuk merencanakan strategi dan alokasi sumber daya yang lebih efektif dalam pelayanan kesehatan di Jawa Barat, serta menyesuaikan pelayanan secara tepat dan efisien sesuai dengan kebutuhan masyarakat di masing-masing wilayah.

Kata kunci : *Universal Health Coverage, BPJS Kesehatan, Segmentasi Daerah, Algoritma K-means, Visualisasi Quantum geographic information system.*

1. PENDAHULUAN

Peningkatan kesehatan global telah menjadi agenda bersama di berbagai negara dan organisasi internasional. Organisasi Kesehatan Dunia (WHO) memperkirakan setidaknya separuh populasi dunia tidak memiliki akses terhadap layanan kesehatan esensial. Laporan WHO dan Bank Dunia tahun 2017 menyebutkan minimal 100 juta orang jatuh miskin akibat harus menanggung biaya kesehatan langsung setiap tahunnya. Untuk mengatasi tantangan ini, konsep *Universal Health Coverage* (UHC) telah mendapatkan momentum global dengan tujuan memastikan setiap individu dan komunitas menerima layanan kesehatan berkualitas tanpa menghadapi kesulitan finansial. Di Indonesia, sistem asuransi kesehatan berperan sebagai salah satu instrumen utama untuk mewujudkan UHC. Program Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan bertindak sebagai penyelenggara Jaminan Kesehatan Nasional (JKN). Dalam rangka meningkatkan akses pelayanan kesehatan yang merata, BPJS Kesehatan menerapkan program JKN dan Kartu Indonesia Sehat (KIS) di Jawa Barat. Meskipun demikian, upaya untuk mengoptimalkan partisipasi masyarakat dan memastikan iuran yang terjangkau bagi seluruh lapisan masyarakat masih perlu menjadi perhatian.

Jawa Barat merupakan salah satu provinsi dengan kepadatan penduduk yang tinggi serta keragaman sosial ekonomi yang signifikan. Penerapan JKN dan Kartu Indonesia Sehat (KIS) oleh BPJS Kesehatan menjadi langkah strategis dalam memberikan akses pelayanan kesehatan yang merata. Namun, upaya untuk mengoptimalkan partisipasi dan memastikan iuran terjangkau bagi seluruh lapisan masyarakat perlu diperhatikan. Penelitian ini bertujuan untuk melakukan segmentasi wilayah di Jawa Barat berdasarkan jumlah warga yang terdaftar sebagai peserta BPJS Kesehatan menggunakan algoritma K-Means.

Metode *K-Means* merupakan salah satu metode *clustering* yang efektif dan banyak digunakan dalam berbagai penelitian sebelumnya. Dalam penelitian ini, algoritma *K-Means* akan diterapkan untuk mengelompokkan wilayah di Jawa Barat berdasarkan kesamaan jumlah warga terdaftar BPJS Kesehatan. Proses optimalisasi jumlah *cluster* akan dilakukan dengan metode *Elbow* dan *Silhouette Coefficient*. Hasil *clustering* akan divisualisasikan menggunakan *Quantum Geographic Information System* (QGIS) untuk memberikan gambaran spasial yang lebih jelas.

Penelitian ini diharapkan dapat memberikan informasi berharga kepada pihak berwenang dan penyelenggara layanan kesehatan di Jawa Barat dalam merencanakan strategi dan alokasi sumber daya yang lebih efektif berdasarkan pemahaman mendalam tentang kebutuhan masyarakat. Dengan demikian,

pelayanan kesehatan dapat disesuaikan secara lebih tepat dan efisien untuk mencapai tujuan UHC di wilayah tersebut.

2. TINJAUAN PUSTAKA

2.1. Sistem Jaminan Sosial Kesehatan

Sistem Jaminan Sosial Kesehatan merupakan sebuah sistem yang bertujuan untuk memberikan perlindungan finansial dan akses layanan kesehatan kepada seluruh penduduk suatu negara[1]. Tujuan utamanya adalah untuk menjadikan kesehatan sebagai hak dasar setiap individu dan memastikan biaya layanan kesehatan tidak menjadi beban berat bagi masyarakat.

2.2. *Data Mining*

Data *mining* adalah tahapan dalam proses penemuan pola atau informasi signifikan dalam *dataset* yang telah dipilih dan dicapai dengan menerapkan berbagai teknik atau metode[2].

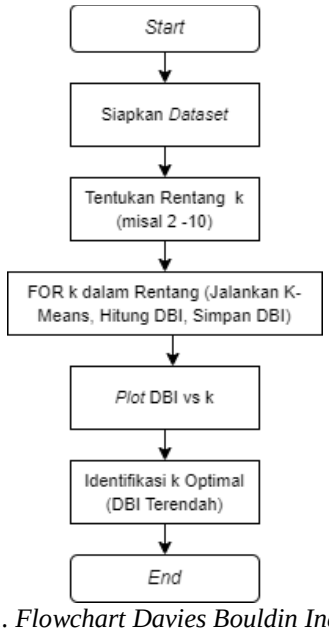
2.3. *Knowledge Discovery in Database* (KDD)

Knowledge Discovery in Database (KDD) merupakan proses untuk menggali informasi pada data yang besar. Dalam hal ini, KDD memiliki keterkaitan dengan *data mining*, di mana data mining adalah bagian dari proses KDD[3]. Tahapan proses KDD meliputi *data selection*, *data proprocessing*, *data transformation*, *data mining* dan *evaluation*.

2.4. *Clustering*

Clustering merupakan proses yang melibatkan pengelompokan titik data ke dalam dua kelompok atau lebih, dengan tujuan agar titik data dalam kelompok yang sama memiliki kesamaan yang lebih besar dibandingkan dengan titik data dalam kelompok yang berbeda[4].

2.5. *Davies Bouldin Index*

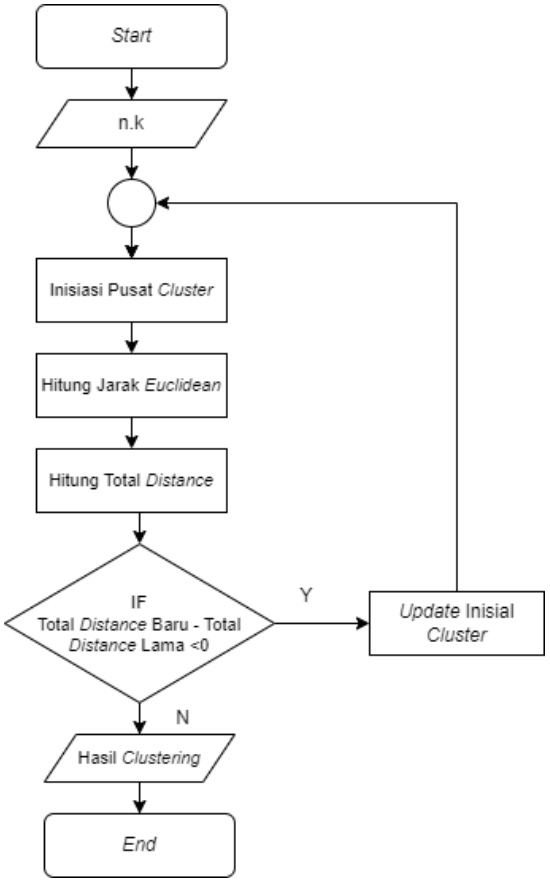


Gambar 1. *Flowchart Davies Bouldin Index* (DBI)

Davies-Bouldin Index (DBI) digunakan untuk mengukur seberapa baik setiap *cluster* terpisah dari *cluster* lainnya. Semakin rendah nilai DBI, semakin baik pemisahan antar cluster, yang menunjukkan kualitas *clustering* yang lebih baik[5]. Oleh karena itu, tujuannya adalah untuk mencari jumlah *cluster* yang menghasilkan nilai DBI terendah, yang mengindikasikan hasil *clustering* yang optimal. Alur pada evaluasi ini dapat dilihat pada gambar 1.

2.6. Algoritma *K-Means*

Algoritma *K-Means* berfokus pada konsep yang relatif sederhana. Langkah awalnya adalah menentukan jumlah *cluster* yang diinginkan, kemudian memilih elemen pertama dalam setiap *cluster* secara acak untuk berperan sebagai titik pusat atau *centroid*[6]. Dalam algoritma *K-Means*, setelah menentukan jumlah *cluster*, titik awal untuk setiap *cluster* diambil dari elemen-elemen yang ada dalam *dataset*. Alur pada algoritma ini dapat dilihat pada gambar 2 sebagai berikut.



Gambar 2. *Flowchart* Algoritma *K-Means*

2.7. Metode *Elbow*

Metode *elbow* merupakan pendekatan yang digunakan dalam analisis *clustering* untuk membantu menentukan jumlah *cluster* yang optimal dalam suatu *dataset*[7]. Alur metode *elbow* yang digunakan dapat dilihat pada gambar 3 sebagai berikut.

Gambar 3. *Flowchart* Metode *Elbow*

2.8. *Sum of Square Error*

Sum of Square Error (SSE) digunakan untuk mengevaluasi tingkat kesalahan atau perbedaan antara data yang telah diperoleh dengan model prediksi yang telah dibuat sebelumnya[8]. Semakin rendah nilai SSE, semakin baik model tersebut dalam melakukan *fitting* atau memprediksi data. Berikut rumus tertentu yang digunakan untuk menghitung *Sum of Square Error* (SSE):

$$SSE = \sum_{k=1}^{k=n} \sum xi \in S_k \parallel X_i - C_k \parallel \quad (1)$$

2.9. *Python*

Python merupakan bahasa pemrograman dengan tipe interpreted language, di mana kode tidak akan diubah menjadi kode yang dapat dipahami oleh komputer sebelum dijalankan, melainkan perubahan tersebut terjadi di *runtime*[9].

2.10. *Google Colaboratory*

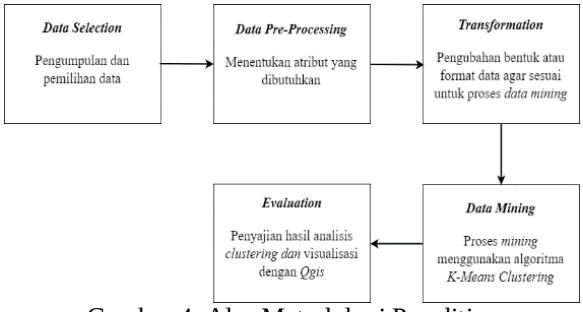
Google colaboratory adalah sebuah proyek yang bertujuan untuk menyebarluaskan pendidikan *machine learning* dan penelitian[10]. *Google colaboratory* merupakan hasil kolaborasi antara *Jupyter* dan *Google*, dimana *Jupyter* sebagai mesin sedangkan *Google* sebagai dokumennya.

2.11. *Quantum Geographic Information System*

QGIS merupakan aplikasi Sistem Informasi Geografis (SIG) yang berguna dalam melakukan analisis dan visualisasi data geografis[11]. QGIS ini memungkinkan pengguna untuk memvisualisasikan data geografis dalam bentuk peta.

3. METODE PENELITIAN

Metodologi penelitian yang digunakan adalah *Knowledge Discovery in Database* (KDD) dengan tahapan proses meliputi *Data Selection*, *Pre-Processsing*, *Transformation*, *Data mining*, dan *Evaluation*. Alur Metodologi KDD dapat dilihat pada gambar 4 sebagai berikut.



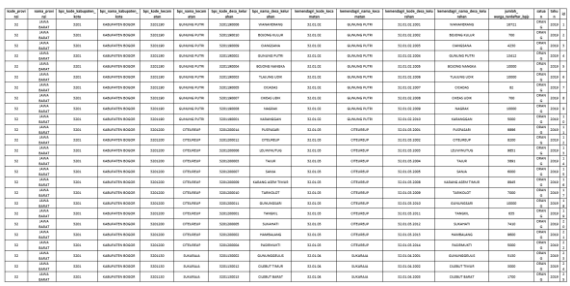
Gambar 4. Alur Metodologi Penelitian

4. HASIL DAN PEMBAHASAN

Hasil utama dari penelitian *data mining* yang telah dilakukan adalah penerapan Algoritma *K-means* dalam melakukan segmentasi wilayah berdasarkan jumlah penduduk yang terdaftar menjadi peserta BPJS Kesehatan di Jawa Barat. Proses *clustering* ini dapat dimanfaatkan sebagai sumber informasi penting bagi pemerintah daerah, pihak BPJS Kesehatan, dan pemangku kepentingan terkait untuk pengambilan keputusan yang lebih baik dalam merencanakan dan meningkatkan layanan kesehatan. Penelitian ini menggunakan metodologi *Knowledge Discovery in Database* (KDD) sebagai metode standar dalam *data mining* untuk mengolah data mentah yang diperoleh dari observasi, dengan tahapan yang meliputi *Data Selection*, *Data Preprocessing*, *Data Transformation*, *Data Mining*, dan *Evaluation* yang akan diuraikan sebagai berikut.

4.1. *Data Selection*

Tahap *data selection* ini merupakan tahap pengumpulan *dataset* awal. Jumlah warga terdaftar menjadi peserta BPJS Kesehatan merupakan data yang dianalisis. Data ini diperoleh dari website opendata.jabarprov.go.id yang dikelola oleh Dinas Pemberdaya Masyarakat dan Desa. Data yang didapatkan disajikan dengan jumlah tupel 26.559 dengan 16 atribut awal dalam bentuk tabel yang ditampilkan pada gambar 5 sebagai berikut



Gambar 5. *Dataset* awal

4.2. *Data Pre-processing*

Data preprocessing dinilai penting karena data dengan kualitas yang tinggi akan menghasilkan *output* dengan kualitas yang tinggi juga. Tahapan *data preprocessing* yang diterapkan terhadap *dataset* meliputi penanganan *missing values* dan *duplicate data.*

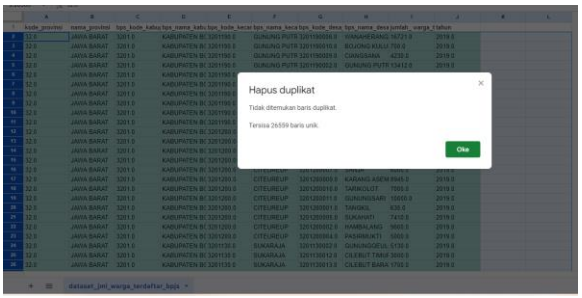
```
[185] # Pengecekan Missing Value

      missing_values = df.isnull().sum()
      print("Jumlah missing value per kolom:")
      print(missing_values)

      Jumlah missing value per kolom:
      kode_provinsi                0
      bps_kode_kabupaten_kota      0
      bps_nama_kabupaten_kota      0
      bps_kode_kecamatan           0
      bps_nama_kecamatan           0
      bps_kode_desa_kelurahan      0
      bps_nama_desa_kelurahan      0
      jumlah_ warga_terdaftar_bpjs 0
      tahun                        0
      dtype: int64
```

Gambar 6. Pengecekan *missing value* menggunakan *google colaboratory*

Berdasarkan gambar 6, tidak ditemukan adanya *missung value* pada data yang akan digunakan. *Tools Microsoft Excel* dan *Google Colaboratory* menjadi *tools* yang digunakan untuk pengecekan *duplicate data*.



Gambar 7. Pengecekan *duplicate data* menggunakan *microsoft excel*

```
[186] # Pengecekan Duplicate Data
      duplicate_rows = df.duplicated().sum()
      print("\nJumlah duplicate data:", duplicate_rows)

      Jumlah duplicate data: 0
```

Gambar 8. Pengecekan *duplicate data* menggunakan *google colaboratory*

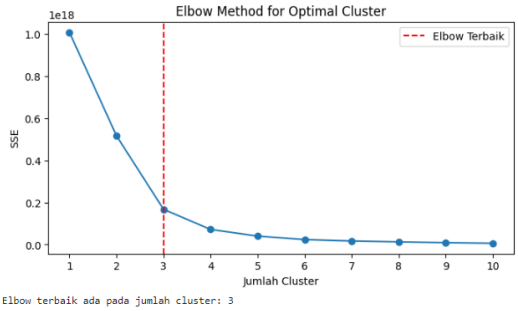
Berdasarkan gambar 7 dan 8 tidak ditemukan adanya *duplicate data* pada data yang akan digunakan.

4.3. *Transformation*

Pada tahap ini dilakukan proses perubahan tipe data kategorikal menjadi numerikal dengan tujuan agar dapat dilakukan proses *clustering k-means*. Metode *K-means* hanya dapat memproses tipe data numerik karena cara kerjanya yang bergantung pada konsep geometris seperti jarak *Euclidean* untuk menentukan kesamaan antara titik-titik data dan *centroid* (pusat *cluster*). Oleh karena itu, setelah melakukan analisis dan eksplorasi terhadap data yang tersedia, semua atribut sudah sesuai dalam bentuk yang diperlukan ke dalam proses data mining sehingga dalam tahapan ini tidak dilakukan proses perubahan tipe data.

4.4. *Data Mining*

Dalam menganalisis pola yang terdapat pada data serta mendapatkan nilai *cluster* optimal, penulis menggunakan metode *elbow* dan Nilai SSE (*Sum of Square Error*) pada tahap *data mining*.



Gambar 9. Grafik Metode *Elbow*

Berdasarkan gambar 9 grafik metode *elbow* menunjukkan bahwa titik siku terbentuk pada *cluster* 3. Kemudian setelah jumlah *cluster* yang optimal diperoleh, tahapan berikutnya adalah implementasi algoritma *k-means* dalam proses *clustering*.

```
[199] # Proses Clustering K-Means dengan jumlah cluster = 3

      kmeans = KMeans(n_clusters=3, random_state= 42)

      # Normalisasi data menggunakan MinMaxScaler
      scaler = MinMaxScaler()
      data = scaler.fit_transform(df_clustering)
      kmeans.fit(data)

      # Mendapatkan centroid dan label
      centroid = kmeans.cluster_centers_
      print(centroid)
      kmeans.labels_

      /usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The defaul
        warnings.warn(
      [[0.         0.09514068 0.09655597 0.09655605 0.02787797 0.87498823]
       [0.         0.09514983 0.09656509 0.09656516 0.01755874 0.125     ]
       [0.         0.09514983 0.09656509 0.09656516 0.02150414 0.5       ]]
      array([1, 1, 1, ..., 0, 0, 0], dtype=int32)
```

Gambar 10. Proses *Clustering K-Means* dan Perolehan Hasil *Centroid* Tiap Atribut

Pada gambar 10, ditampilkan nilai pusat atau *centroid* dari masing-masing cluster secara acak untuk mengelompokkan *cluster* yang paling mirip.

Tabel 1. Perhitungan nilai SSE

Jumlah *Cluster*	**Nilai SSE**
1	1.00783
2	5.16576
3	1.67409
4	7.18240

Berdasarkan hasil perhitungan nilai SSE pada tabel 1 dan grafik metode *elbow* dengan nilai SSE paling rendah dan titik siku berada C=3 diperoleh jumlah *cluster* optimal adalah 3 *cluster*.

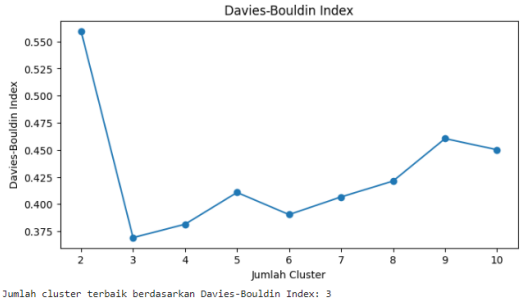
4.5. *Evaluation*

Untuk melakukan evaluasi *cluster*, peneliti menggunakan parameter *Davies Bouldin Index* (DBI) untuk mengukur rasio jarak antar *cluster* terhadap jarak dalam *cluster*. Kualitas *cluster* akan semakin baik jika nilai DBI yang dihasilkan semakin kecil.

Tabel 2. Hasil Pengujian *Davies Bouldin Index* (DBI)

Jumlah *Cluster*	Hasil DBI
1	-
2	0.559471560095628
3	0.36915605275672103
4	0.3814990903091133

Berdasarkan tabel 2, dihasilkan perhitungan yang menunjukkan bahwa nilai DBI terendah diperoleh pada 3 *cluster* dengan nilai 0,36915605275672103. Hal ini mengindikasikan bahwa pada jumlah *cluster* 3, objek dalam *cluster* memiliki kesamaan yang tinggi dan perbedaan yang signifikan dengan objek di *cluster* lain, sehingga menghasilkan kualitas *cluster* yang baik.



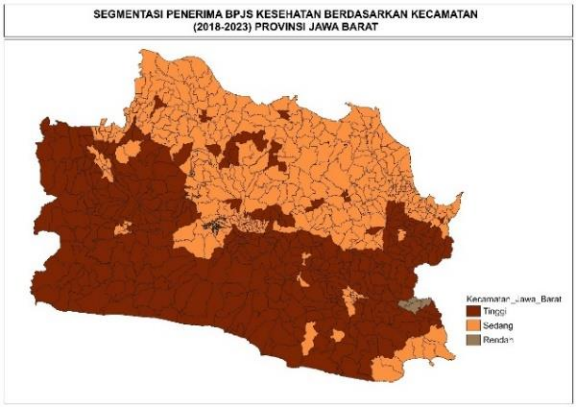
Gambar 11. Grafik *Davies Bouldin Index* (DBI)

Pada gambar 11, memampilkan grafik DBI yang dihasilkan adalah K=3. Selanjutnya *cluster-labeling* dilakukan berdasarkan ciri khas dari setiap *cluster*, menghasilkan tiga label: kategori tertinggi, menengah dan terendah yang disajikan pada tabel 2.

Tabel *2. Hasil Cluster*

Cluster	Jumlah Berdasarkan Desa	Jumlah Berdasarkan Kecamatan	Kategori
0	14060	501	Tertinggi
1	12419	222	Menengah
2	80	4	Rendah

Selanjutnya, hasil pengklasteran akan dilakukan visualisasi pemetaan menggunakan *tools* QGIS.



Gambar *9. Hasil Visualisasi*

Pada Gambar 9, warna merah untuk daerah dengan jumlah peserta BPJS Kesehatan tertinggi, warna *orange* untuk daerah dengan jumlah perserta BPJS Kesehatan menengah, dan warna coklat untuk daerah dengan jumlah peserta BPJS terendah.

5. KESIMPULAN DAN SARAN

Penerapan teknik *data mining clustering* menggunakan algoritma *k-means* untuk pengelompokan wilayah di Jawa Barat berdasarkan jumlah warga terdaftar BPJS pada *dataset* tahun 2019-2023 menghasilkan 3 *cluster*. *Cluster* 0 dengan 501 kecamatan memiliki jumlah peserta BPJS Kesehatan tinggi, *Cluster* 1 dengan 222 kecamatan memiliki jumlah peserta BPJS Kesehatan menengah, dan *Cluster* 2 dengan 4 kecamatan memiliki jumlah peserta BPJS Kesehatan rendah. Evaluasi kinerja algoritma menunjukkan *cluster* 3 memiliki *Sum of Square Error* (SSE) terkecil sebesar 1,67409 dan nilai *Davies Bouldin Index* (DBI) terendah 0,36915605275672103, sehingga menjadi *cluster* terbaik. Visualisasi dengan *Quantum Geographic Information System* menunjukkan warna merah untuk *cluster* 0, warna *orange* untuk *cluster* 1, dan warna coklat untuk *cluster* 2. Untuk penelitian selanjutnya, disarankan untuk menggunakan *dataset* yang lebih banyak atau terbaru, mencoba algoritma atau metode *clustering* lain, menggunakan *tools* data *mining* yang lebih akurat, serta menggunakan parameter evaluasi hasil *clustering* selain *Davies Bouldin Index* (DBI) untuk mendapatkan hasil yang lebih komprehensif.

DAFTAR PUSTAKA

[1] N. Azwanti and N. E. Putria, "Penerapan Algoritma K-Means Untuk Pemetaan Penerimaan Bantuan Kesejahteraan Masyarakat di Kota Batam," *Semin. Nas. Ilmu Sos. dan Teknol.*, no. September, pp. 250–256, 2023.

[2] R. B. B. Sumantri and E. Utami, "Penentuan Status Tahapan Keluarga Sejahtera Kecamatan Sidareja Menggunakan Teknik Data Mining," *Respati*, vol. 15, no. 3, p. 71, 2020, doi: 10.35842/jtir.v15i3.375.

[3] Y. M. Arianti Jati, Agus Sunandar, "DARAH DENGAN ALGORITMA NAÏVE BAYES DAN

METODE KNOWLEGDE DISCOVERY IN DATABASE (KDD) (Studi Kasus : PMI Kabupaten . Karawang),” vol. 2, no. 1, pp. 1–9, 2023.

[4] K. K. Munawar and A. I. Purnamasari, “IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING PADA KLASTERISASI KASUS HIV DI JAWA BARAT,” vol. 7, no. 2, 2023.

[5] E. Muningsih, I. Maryani, and V. R. Handayani, “Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa,” *J. Sains dan Manaj.*, vol. 9, no. 1, pp. 95–100, 2021, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/view/10428/4839.

[6] A. F. Sitepu, “Penerapan Data Mining Pengelompokan Data Pasien Berdasrkan Jenis Penyakit Menggunakan Metode Clustering (Studi Kasus Klinik Mitra Nd),” *J. Sist. Inf. Kaputama*, vol. 6, no. 2, pp. 192–207, 2022, doi: 10.59697/jsik.v6i2.169.

[7] A. Winarta and W. J. Kurniawan, “Optimasi Cluster K-means Menggunakan Metode Elbow pada Data Pengguna Narkoba dengan Pemrograman Python,” *J. Tek. Inform. Kaputama*, vol. 5, no. 1, pp. 113–119, 2021.

[8] D. J. Kadhim, M. H. Saleh, and S. J. Abou-Loukh, “Evaluation of massive multiple-input multiple-output communication performance under a proposed improved minimum mean squared error precoding,” *IAES Int. J. Artif. Intell.*, vol. 12, no. 2, pp. 984–994, 2023, doi: 10.11591/ijai.v12.i2.pp984-994.

[9] Y. S. Sendi Algifari Risnawan, “Implementasi Website Berita Online Menggunakan Metode Crawling Data Dengan Bahasa Pemrograman Python,” *Se*, vol. 13, no. 01, pp. 31–34, 2021.

[10] A. L. F. Edi Supriyadi, Jarnawi Afgani Dahlan, Dadang Juandi, Rani Sugiarni, Sarah Inayah, Samsul Pahmi, Ratu Sarah Fauziah Iskandar, Mahmudin, “ANALISIS SENTIMEN BERDASARKAN ULASAN PENGGUNA APLIKASI MYPERTAMINA PADA GOOGLE PLAYSTORE MENGGUNAKAN METODE NAÏVE BAYES,” *J. Ilm. Tek. dan Ilmu Komput.*, vol. 2, no. 3, pp. 100–108, 2023.

[11] I. Zulfa, R. Septima, M. Handri, I. Zulfida, and L. Suryati, “Pemetaan Wilayah Persebaran Padi dan Kopi dengan Quantum,” vol. 3, no. 6, pp. 459–468, 2023.