BIG DATA CHALLENGE (BDC)
SATRIA DATA 2022
Universitas Islam Indonesia

Klasterisasi Daerah berdasarkan Analisis Sentimen Twitter dan Data Klaim BPJS sebagai Upaya Peningkatan Layanan BPJS di Indonesia menggunakan K-Means *Clustering*

Klasterisasi Klaim dan Sentimen BPJS Kesehatan di Indonesia

BDC_SD20220000074

1. Pendahuluan

Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan merupakan badan nasional yang bertujuan untuk menyelenggarakan program jaminan kesehatan di atas prinsip asuransi sosial dan ekuitas sebagaimana merujuk pada UU No. 40 Tahun 2004. Seluruh warga Indonesia wajib mendaftar sebagai peserta program BPJS yang dibagi menjadi beberapa kategori kepesertaan: 1) Pekerja Penerima Upah, 2) PD Pemda, 3) Pekerja Bukan Penerima Upah (PBPU) dan Bukan Pekerja (BP), dan 4) Penerima Bantuan Iuran Jaminan Kesehatan (PBI JK). Manfaat yang didapatkan sebagai peserta program Jaminan Kesehatan Nasional (JKN) BPJS beragam, seperti pelayanan kesehatan tingkat pertama berupa pelayanan kesehatan yang bersifat non spesialistik, pelayanan kebidanan, ibu, bayi dan balita, rawat inap non intensif dan intensif, serta manfaat lainnya. Namun, dengan berbagai manfaat yang ditawarkan oleh BPJS Kesehatan, terdapat banyak keluhan dari masyarakat mengenai penyelenggaraan program ini, di antaranya adalah permasalahan kualitas layanan, penunggakan iuran, keterlambatan klaim, bahkan penolakan layanan peserta BPJS oleh rumah sakit terkait. Hal tersebut tentu menjadi catatan penting bagi BPJS untuk dapat meningkatkan kualitas layanan dan pengawasannya agar dapat lebih prima dalam melayani masyarakat.

Text mining merupakan proses untuk mencari informasi baru dari sekumpulan data dalam bentuk dokumen menggunakan teknik-teknik tertentu sehingga menghasilkan insight baru. *Text mining* berbeda dengan *data mining* dikarenakan *text mining*

berhubungan khusus dengan data yang tidak terstruktur atau semi-terstruktur berupa teks. Di antara penerapan *text mining* adalah *scraping* data cuitan/*tweet* di Twitter untuk dilakukan keperluan analisis sentimen. *Scraping* dilakukan menggunakan *library* bantuan sebagai jembatan terhadap API yang telah disediakan oleh Twitter atau dengan *scraping* melalui *front-end* Twitter menggunakan *keyword* tertentu. Data cuitan Twitter ini kemudian dapat diolah dan dilakukan analisis untuk menghasilkan insight baru, misalnya: *Social Network Analysis* (SNA) atau analisis sentimen. Pada setiap cuitan di Twitter memiliki *tag* atau atribut, seperti: tanggal cuitan, lokasi, user id, dan lainnya. Atribut tersebut dapat dimanfaatkan sebagai data pendukung analisis.

K-means clustering merupakan algoritma *machine learning* yang dapat mengelompokkan kumpulan data ke dalam klaster-klaster. Algoritma ini termasuk ke dalam jenis *unsupervised machine learning* yang mampu mengkategorikan data yang memiliki kesamaan atau kemiripan dengan data lainnya ke dalam satu kelompok.

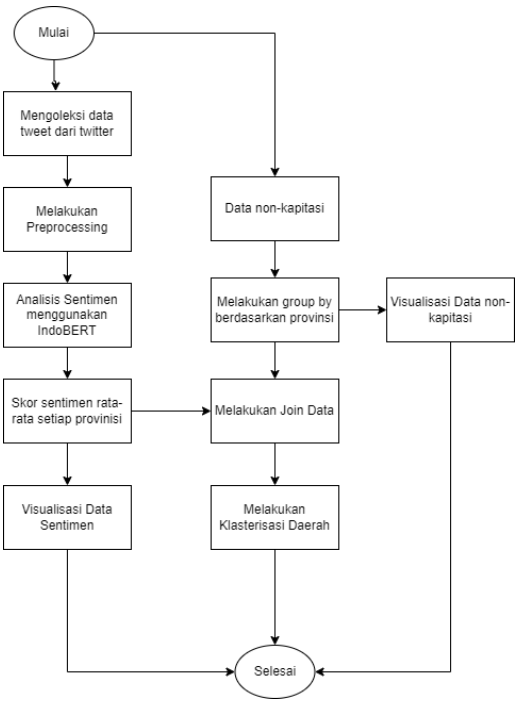
Adapun tujuan dari kegiatan ini yaitu,

- Melihat sebaran karakteristik data non-kapitasi BPJS di setiap provinsi di Indonesia
- Melihat sebaran sentimen pengguna Twitter di setiap provinsi di Indonesia
- Menentukan klaster-klaster daerah tentang pelayanan BPJS di Indonesia

dengan batasan-batasan yakni data yang digunakan berasal dari rentang waktu 1 tahun terakhir, dari tahun 2021 hingga 2022, serta sumber data pendukung yang diperoleh dari proses *scraping* media sosial Twitter dengan bantuan *library* pemroses. Hasil *scraping* kemudian akan dicari sentimennya dengan melakukan *sentiment analysis* untuk mengetahui nilai sentimen pada setiap cuitan yang ada. Selain itu, dilakukan pengklasteran pada data non-kapitasi untuk dilihat karakteristiknya di setiap daerah. Hasil analisis sentimen dan klasterisasi diharapkan dapat dijadikan sebagai informasi daerah yang menjadi prioritas utama perbaikan pelayanan BPJS.

2. Metodologi

Gambar di bawah ini menjelaskan tentang diagram alir dari penelitian sederhana yang dilakukan. Diagram alir tersebut berisi langkah-langkah dari metodologi yang digunakan pada penelitian sederhana ini.



Gambar 2.1 Diagram Alir Penelitian

2.1. Mengoleksi Data *Tweet* dari Twitter

Twitter digunakan sebagai sumber data tambahan untuk memperkuat analisis yang dilakukan. Data *tweet* digunakan dikarenakan cukup merepresentasikan opini dari masyarakat serta dapat diakses secara publik dengan bantuan *library* melalui teknik *scraping*. Untuk teknik *scraping* sendiri digunakan bantuan *library* snscrape. Snscrape digunakan karena kapabilitasnya yang tidak memerlukan API dari Twitter, yang memiliki batasan kuota dalam melakukan *scraping*.

Scraping sendiri dilakukan untuk memperoleh *tweet* masyarakat dari tiap provinsi yang dilihat dari 1 tahun terakhir. Adapun *query* yang digunakan untuk melakukan *scraping* pada provinsi tertentu terlihat pada potongan kode di bawah berikut.

```
query = 'bpjs near:"' + city + '" within:20km
```

```
since:2021-01-01 until:2022-10-10 lang:id'
df_temp =
pd.DataFrame(itertools.islice(sntwitter.TwitterSearchScra
per(
      query).get_items(), 200))[['date', 'content']]
```

Pada variabel *query*, kata “bpjs” merupakan kata kunci dari pencarian *tweet*, yang digunakan untuk melihat *tweet* yang berhubungan dengan BPJS sebagai layanan yang menjadi fokus utama. Variabel *city* dapat diisi dengan nama kota tiap provinsi yang ingin dicek. Sebagai contoh apabila ingin melihat *tweet* di provinsi Riau, maka variabel city dapat diganti dengan “Pekanbaru” yang merupakan ibukota dari Pekanbaru itu sendiri. Radius pencarian dari titik pusat dapat diatur dengan mengubah variabel “within”. Sedangkan untuk variabel “since” dan “until” digunakan untuk mengatur rentang waktu pencarian *tweet*, di mana pada kode di atas waktu yang dicari adalah sejak 1 Januari 2021 hingga 10 Oktober 2022.

Masing-masing *tweet* per daerah dicari dengan limit *tweet* per daerah sebanyak 200 *tweets*. Dari hasil pencarian, diperoleh bahwa tidak seluruh daerah mencapai 200 *tweets* yang membahas mengenai BPJS, umumnya pada daerah yang tidak terlalu padat seperti daerah di luar Pulau Jawa.

2.2. Melakukan *Preprocessing Data*

Setelah memperoleh *tweet* dari setiap provinsi mengenai pelayanan BPJS, selanjutnya dilakukan *preprocessing* data terlebih dahulu untuk membersihkan *tweet* agar lebih mudah dimengerti oleh model *sentiment analysis* nantinya. Tahap *preprocessing* sendiri terdiri dari menghapus tanda baca, menghapus angka, menghapus *stopwords* dalam Bahasa Indonesia, hingga melakukan *stemming* untuk mengubah kata berimbuhan ke dalam bentuk dasarnya. Hasil dari *preprocessing* data *tweet* ini yaitu *tweet* yang lebih ‘bersih’ dan dapat dimasukkan ke dalam model *sentiment analysis*.

2.3. Analisis Sentimen menggunakan IndoBERT

Tweet yang sudah dipraproses sebelumnya akan dimasukkan ke dalam model *pretrained* dari IndoBERT. Sesuai namanya, IndoBERT sendiri adalah versi

Bahasa Indonesia dari model BERT yang dapat digunakan untuk berbagai tujuan, salah satunya yaitu untuk *sentiment analysis*. BERT sendiri merupakan singkatan dari Bidirectional Encoder Representations from Transformers, di mana berbeda dengan model pemrosesan bahasa lainnya, BERT didesain untuk melatih teks tanpa label menghasilkan model yang dapat digunakan pada berbagai macam tugas (Devlin et al. 2019). Dari model *pre trained* ini, kita dapat menggunakan baik model maupun tokenizer yang sudah disediakan untuk memprediksi sentimen dari masing-masing *tweet* dengan akurasi yang tinggi.

Dari hasil prediksi sentimen, ditemukan bahwa ada sekitar 900 *tweet* dengan sentimen negatif, 300 *tweet* dengan sentimen positif, dan 1900 *tweet* dengan sentimen netral, dengan total sejumlah 3141 *tweet*. Dari tiap sentimen yang sudah dideteksi dari tiap *tweet*, langkah selanjutnya adalah untuk menggabungkan keseluruhan sentimen tersebut per provinsi untuk melihat rata-rata sentimen tiap provinsi. Dari sini, diperoleh bahwa daerah dengan rata-rata sentimen di bawah <0.5 menandakan sentimen yang diperoleh dari *tweet* yang ada mengarah ke negatif, sedangkan apabila rata-rata sentimen bernilai >0.5 maka berarti sebaliknya, dengan nilai 0.5 menandakan rata-rata sentimen netral.

2.4. Melakukan *Group By* Provinsi pada Data non-kapitasi

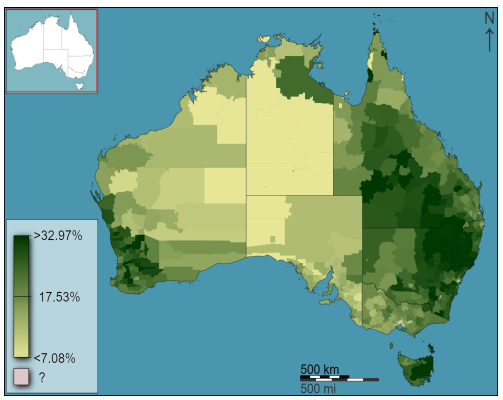
Data non-kapitasi yang diberikan oleh pihak Satria Data merupakan data besaran pembayaran klaim oleh BPJS Kesehatan kepada Fasilitas Kesehatan Tingkat Pertama berdasarkan jenis dan jumlah pelayanan kesehatan yang diberikan. Data tersebut masih berupa data baris dari setiap kunjungan atau klaim di berbagai daerah. Kemudian penulis melakukan pengelompokan untuk setiap provinsi yang menunjukan karakteristik dari klaim dan penggunaan BPJS. Variabel-variabel yang dibentuk dari data non-kapitasi untuk setiap provinsi adalah sebagai berikut :

- Jumlah Peserta
- Jumlah KK
- Rata-rata Bobot
- Jumlah Kunjungan
- Rata-rata Lama Rawat

- Jumlah Kota
- [Jumlah Swasta, Pemerintahan Kab/Kota, dll] untuk kepemilikan faskes
- [Jumlah laboratorium, puskesmas, klinik pratama dll] untuk jenis faskes
- [Jumlah laboratorium, rawat inap, dll] untuk tipe faskes
- [Jumlah promotif, rjtp, dll] untuk tingkat pelayanan
- [Jumlah ppu, pbu, dll] untuk segmentasi peserta
- [Jumlah E119, I10, dll] untuk diagnosa penyakit
- [Jumlah Gula Darah Puasa (GDP) - PRB/Prolanis, Paket Persalinan per Vaginam normal (oleh Bidan) dll] untuk nama tindakan
- [Jumlah biaya klaim, rata-rata biaya klaim] untuk biaya klaim

2.5. Visualisasi Data non-kapitasi dan Sentimen

Visualisasi dilakukan menggunakan *Choropleth Map*. *Choropleth Map* adalah peta tematik yang digunakan untuk merepresentasikan data statistik dengan menggunakan teknik simbologi pemetaan warna. Ini menampilkan unit pencacahan, atau membagi wilayah atau wilayah geografis yang diwarnai, diarsir, atau berpola dalam hubungannya dengan variabel data. Untuk menunjukkan variasi atau pola di seluruh lokasi yang ditampilkan, peta choropleth menyediakan cara untuk memvisualisasikan nilai di wilayah geografis (Pedriquez, 2022). Pada visualisasi Data non-kapitasi, penulis merepresentasikan beberapa data yaitu sebaran data Jumlah Peserta, Jumlah Kunjungan, Rata-rata lama rawat, serta Rata-rata biaya klaim. Selanjutnya penulis juga akan merepresentasikan data rata-rata sentimen untuk setiap provinsi.



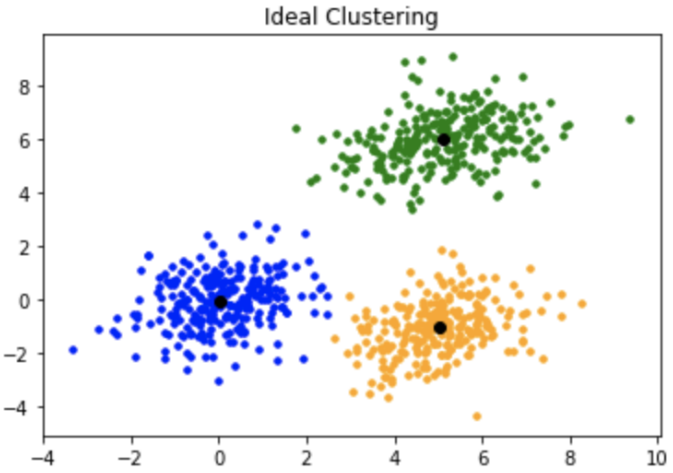
Gambar 2.2 Contoh *Choropleth Map*

2.6. Melakukan *Join* pada Data Sentimen dan Data non-kapitasi

Setelah penulis mengelompokan data non-kapitasi dan data hasil analisis sentimen berdasarkan provinsi. Penulis akan menggabungkan atau melakukan join pada kedua data tersebut berdasarkan nama provinsi yang sama.

2.7. Melakukan Klasterisasi Daerah menggunakan K-Means Clustering

Penulis akan melakukan klasterisasi terhadap data hasil join pada tahap 2.6 menggunakan K-Means. K-Means clustering merupakan algoritma yang membutuhkan input sebanyak k yang membagi n objek ke dalam k cluster sehingga tingkat kemiripan antar anggota dalam satu cluster tinggi sedangkan tingkat kemiripan dengan anggota pada cluster lain sangat rendah (Dubey A K Gupta U dan Jain S 2018). K-Means pada penelitian ini akan melakukan klasterisasi untuk melihat 3 klaster yang berbeda pada data tersebut.



Gambar 2.3 Contoh *Clustering*

3. Pembahasan

3.1. Eksplorasi karakteristik data non-kapitasi

3.1.1. Jumlah Peserta

Gambar 3.1 Jumlah Peserta Berdasarkan Wilayah

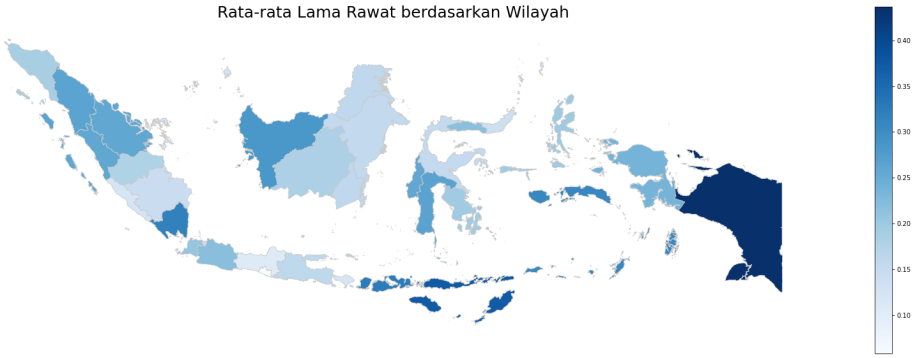
Berdasarkan Gambar 3.1 dapat dilihat bahwa jumlah peserta BPJS masih berpusat di pulau jawa saja seperti provinsi Jawa Barat, Jawa Tengah, dan Jawa Timur. Sementara untuk daerah-daerah di luar pulau jawa jumlah peserta BPJS yang menggunakan BPJSnya masih sedikit. Provinsi dengan jumlah peserta BPJS terbanyak adalah Jawa Tengah

3.1.2. Jumlah Kunjungan

Gambar 3.2 Jumlah Kunjungan Berdasarkan Wilayah

Berdasarkan Gambar 3.2 Jumlah kunjungan pasien yang menggunakan BPJS juga masih terpusat di daerah jawa saja seperti Jawa Barat, Jawa Tengah, dan Jawa Timur. Provinsi dengan jumlah kunjungan pasien menggunakan BPJS terbanyak adalah Jawa Tengah.

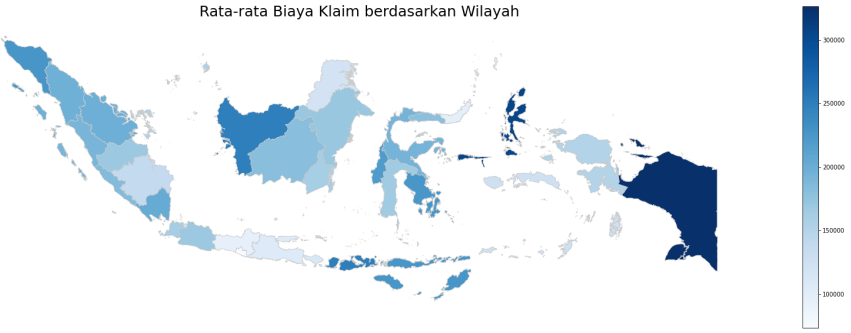
3.1.3. Rata-rata Lama Rawat



Gambar 3.3 Rata-rata Lama Rawat Berdasarkan Wilayah

Lama rawat dihitung dari tanggal datang dan tanggal pulang pada fasilitas kesehatan. Provinsi dengan rata-rata lama rawat terbesar adalah Papua, kemudian Nusa Tenggara.

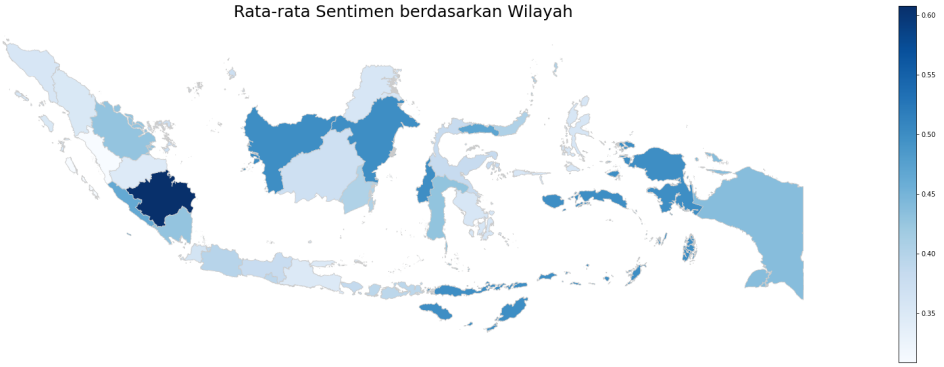
3.1.4. Rata-rata Biaya Klaim



Gambar 3.4 Rata-rata Biaya Klaim Berdasarkan Wilayah

Gambar 3.4 Menunjukan bahwa Provinsi Papua memiliki rata-rata biaya klaim tertinggi setiap kunjungannya. Kemudian ada Maluku Utara dan Kalimantan Barat.

3.2. Eksplorasi Hasil Analisis Sentimen



Gambar 3.5 Rata-rata Sentimen Klaim Berdasarkan Wilayah

Gambar 3.5 menunjukan rata-rata skor sentimen berdasarkan provinsi di Indonesia. Rata-rata sentimen 1 menunjukan positif, 0.5 menunjukan netral, dan 0 menunjukan negatif. Sentimen di setiap provinsi di Indonesia beragam. Untuk Pulau Jawa dan Bali yang merupakan daerah dengan jumlah peserta BPJS dan jumlah kunjungan terbanyak, memiliki sentimen yang cenderung negatif. Di Pulau Sumatera bagian barat dan utara juga cenderung memiliki sentimen negatif. Riau dan Lampung memiliki sentimen netral sementara Sumatera selatan memiliki sentimen positif. Beralih ke Pulau Kalimantan, dapat dilihat bahwa Provinsi Kalimantan Barat dan Kalimantan Utara memiliki sentimen yang cenderung positif sementara Kalimantan Tengah dan Kalimantan Selatan memiliki sentimen yang cenderung negatif. Di Pulau Sulawesi dan Provinsi Maluku Utara cenderung memiliki sentimen netral menuju negatif, Papua Barat dan Maluku cenderung positif, Papua cenderung netral. Nusa Tenggara Barat dan Timur memiliki sentimen yang cenderung positif.

3.3. Klasterisasi Daerah

Gambar 3.6 Hasil Klasterisasi Daerah

Dengan menggunakan Algoritma *K-Means Clustering* penulis membagi daerah menjadi 3 klaster berdasarkan data non-kapitasi dan data sentimen. Dapat dilihat pada Gambar 3.6 terbentuk 3 klaster yang dapat digunakan oleh BPJS untuk melakukan segmentasi daerah dalam pelayanan.

- Klaster 2 yaitu Provinsi Aceh, Sumatera Utara, Lampung, Banten, Kalimantan Barat, Nusa Tenggara, dan Sulawesi Selatan.
- Klaster 1 yaitu Provinsi Jawa Barat, Jawa Timur, Jawa Tengah
- Klaster 0 yaitu Pulau Sumatera kecuali Aceh, Sumatera Utara, Lampung, DI Yogyakarta, Pulau Kalimantan kecuali Kalimantan Barat, Pulau Sulawesi Kecuali Sulawesi Selatan, Maluku Utara, Maluku, serta seluruh Provinsi yang ada di Pulau Papua

4. Kesimpulan

- Jumlah peserta BPJS dan jumlah pasien yang menggunakan BPJS masih terpusat di Pulau Jawa saja. Hal ini tentunya menjadi sinyal bagi pihak BPJS untuk melakukan penyuluhan pada daerah-daerah di luar Pulau Jawa khususnya agar seluruh masyarakat Indonesia dapat menikmati dan terlindungi oleh fasilitas asuransi yang ditawarkan oleh pihak BPJS.
- Rata-rata lama rawat pada pasien pengguna BPJS paling besar ada di Provinsi Papua, kemudian Maluku dan kemudian Nusa Tenggara. Hal tersebut menunjukan bahwa banyak penyakit yang memerlukan rawat inap disana.

- Provinsi Papua, Maluku, dan Kalimantan Barat merupakan Provinsi dengan rata-rata biaya klaim terbesar. Di Provinsi tersebut banyak pengobatan yang memerlukan biaya mahal.
- Analisis sentimen menunjukan bahwa pada Pulau Jawa dan Bali dimana daerah tersebut merupakan daerah dengan jumlah peserta dan kunjungan pasien BPJS terbanyak, serta daerah Sumatera bagian barat dan utara, Kalimantan Tengah, dan sebagian besar wilayah Sulawesi dan Maluku Utara memiliki sentimen negatif. Daerah tersebut seharusnya menjadi konsentrasi utama dan prioritas pihak BPJS untuk meningkatkan layanannya agar opini publik akan BPJS memiliki sentimen positif.
- Terdapat 3 klaster yang dibentuk dari proses klasterisasi yang dapat digunakan untuk melakukan segmentasi daerah dalam pelayanan yang dilakukan oleh pihak BPJS.
    - Klaster 2 yaitu Provinsi Aceh, Sumatera Utara, Lampung, Banten, Kalimantan Barat, Nusa Tenggara, dan Sulawesi Selatan.
    - Klaster 1 yaitu Provinsi Jawa Barat, Jawa Timur, Jawa Tengah
    - Klaster 0 yaitu Pulau Sumatera kecuali Aceh, Sumatera Utara, Lampung, DI Yogyakarta, Pulau Kalimantan kecuali Kalimantan Barat, Pulau Sulawesi Kecuali Sulawesi Selatan, Maluku Utara, Maluku, serta seluruh Provinsi yang ada di Pulau Papua.

5. Referensi

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

Dubey, A. K., Gupta, U., & Jain, S. (2018). Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data. International Journal on Advanced Science, Engineering and Information Technology, 8(1), 18. https://doi.org/10.18517/ijaseit.8.1.3490

Pedriquez, D. (2022, May 17). What is a choropleth map and how to create one. Venngage. Retrieved November 11, 2022, from https://venngage.com/blog/choropleth-map/

Susanto, Hery. (2021). Ombudsman RI Respons Banyaknya Keluhan Soal BPJS Kesehatan. Ombudsman. Diakses pada 12 November 2022. https://ombudsman.go.id/news/r/ombudsman-ri-respons-banyaknya-keluhan-soal-bpjs-kesehatan

Ming-Syan Chen, Jiawei Han and P. S. Yu, "Data mining: an overview from a database perspective," in IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883, Dec. 1996, doi: 10.1109/69.553155.

Orleans, Bryan, et. al. (2022). Clustering Algoritma (K-Means). School of Information Systems BINUS. Diakses pada 12 November 2022. https://sis.binus.ac.id/2022/01/31/clustering-algoritma-k-means/