JURNAL SAINTIKA UNPAM : Jurnal Sains dan Matematika Unpam Vol. 5, No. 2 (2022)
p-ISSN 2621-7856 e-ISSN 2655-7312 6th article, page 63 - 71

ANALISIS CLUSTER *K-MEANS* TENAGA KESEHATAN DI PROVINSI BANTEN

K-MEANS CLUSTER ANALYSIS OF HEALTH WORKERS IN BANTEN PROVINCE

Imamatun Nisa[1)*], Choirul Basir[2)]

[1,2)] Program Studi Matematika, Universitas Pamulang, Tangerang Selatan, Banten
[*]Email: dosen02278@unpam.ac.id

ABSTRACT

Health workers in Indonesia play a very important role in everyday life, both in districts and cities. This can attract research to find out or want to group districts and cities based on data on the number of health workers in Banten Province, which consists of four districts and four cities. The object of this research is the recording of health workers in 2020. The related variables are doctors, dentists, nurses, midwives and pharmacists in health workers in Banten Province. Data collection is obtained from data on health workers. Data processing using clustering and data mining techniques. It was concluded that based on the results of clustering with the K-Means method, it was concluded that there were only three clusters, namely Tangerang district, Tangerang city and South Tangerang city and the rest were included in the zero cluster. The center of the second cluster is an area of high similarity the resulting cluster consists of two clusters namely cluster 0 and cluster 1.

Keywords: Clustering, Data Mining, Health workers, K-Means, Python.

ABSTRAK

Tenaga kesehatan di Indonesia ini sangatlah berperan dalam kehidupan sehari-hari, baik dalam kabupaten maupun kota. Hal ini dapat menarik penelitian untuk mengetahui atau ingin mengelompokkan kabupaten dan kota berdasarkan data jumlah tenaga kesehatan di Provinsi Banten, yang terdiri dari empat kabupaten dan empat kota. Objek penelitian ini dalam pencatatan tenaga kesehatan di tahun 2020. Variabel terkait adalah dokter, dokter gigi, perawat, bidan dan tenaga kefarmasian dalam tenaga kesehatan di Provinsi Banten. Pengumpulan data didapat dari data tenaga kesehatan. Pengolahan data menggunakan clustering dan teknik data mining. Disimpulkan bahwa berdasarkan hasil *clustering* dengan metode *K-Means* maka didapatkan kesimpulan bahwa kelompok cluster satu hanya tiga yaitu kabupaten Tangerang, kota Tangerang dan kota Tangerang Selatan dan selebihnya masuk pada cluster ke nol. Pusat *cluster* ke dua merupakan daerah tingkat kemiripan tinggi, *cluster* yang dihasilkan ada dua *cluster* yaitu *cluster* 0 dan *cluster* 1.

Kata kunci: *Clustering*, Data Mining, Tenaga kesehatan, *K-Means*, Python.

1. PENDAHULUAN

Provinsi Banten merupakan provinsi dengan letak yang strategis karena berbatasan dengan Jawa Barat, ibukota negara DKI Jakarta, dan pulau Sumatera. Perkembangan provinsi Banten akan menjadi penyangga kemajuan wilayah sekitarnya

sehingga diperlukan strategi dalam penataan sumber daya alam dan sumber daya manusianya. Dengan wilayah yang cukup luas dan terdiri dari delapan kota/kabupaten, provinsi ini diharapkan dapat lebih cepat berkembang kemajuannya.

Jumlah dalam tenaga kesehatan di provinsi Banten sebanyak 7.067 dokter, 817 dokter gigi, 14.321 perawat, 7.259 bidan dan 2.322 farmasi. Dalam metode ini akan menggunakan metode data mining yang berperan untuk melihat sumber basis data yang besar dalam pencarian informasi yang berharga sehingga akan didapatkan model data dan nilai-nilainya pada sekumpulan data tersebut. Selain dari pengguna data mining yang akan dilakukan menggunakan teknik *clustering,* dimana teknik penggunaan *clustering* ini dengan melakukan pengelompokkan data jumlah tenaga kesehatan dari beberapa kabupaten/kota berdasarkan jarak minimum setiap data ke *cluster*. Penelitian terkait dengan klaster kota/kabupaten sebagai strategi pengelompokkan kota/kabupaten di Provinsi Banten. Diharapkan dapat memberikan gambaran tentang permasalahan yang ada, efektifitas hubungan terkait yaitu jumlah nakes yang ada. Dengan demikian dapat membantu memberikan informasi terkait jumlah tenaga kesehatan di provinsi Banten.

2. METODOLOGI

Metodologi yang akan digunakan akan menggunakan metode *clustering K-Means* dalam mengelompokkan sekumpulan objek data. Dimana terdapat jumlah tenaga kesehatan di kabupaten atau kota provinsi Banten memiliki tenaga kesehatan yang memadai, hal ini penulis tertarik untuk mengelompokkan data tersebut kedalam metode *clustering K-Means*. *Clustering* ini dilakukan untuk mengelompokan objek yang awalnya menyebar kedalam *cluster* yang diharapkan berdasarkan adanya kemiripan satu dengan yang lainnya. Dalam membagi pengamatan menjadi kelompok, pengamatan tersebut harus dilakukan sedemikian rupa hingga tiap unsur-unsur meminimalkan ragam dalam *cluster* (Rahmawati, Rika, Fatayat, 2020).

2.1. *Clustering*

Pengelompokan sekumpulan objek yang mempunyai kemiripan karakteristik dalam satu kelompok yang sama dari sekumpulan data merupakan bentuk *Clustering*. *Clustering* dalam *machine learning* termasuk dalam kategori *unsupervised learning*, dimana data yang akan akan dikelompokkan belum memiliki label ataupun pola. Pada proses *clustering* ini akan ditentukan terlebih dahulu jumlah *cluster* yang akan dibentuk dari sekumpulan data dengan menggunakan salah satu metode yang tepat dalam

menentukan jumlah *cluster* yang ideal. Setelah proses penentuan jumlah *cluster* ditentukan, selanjutnya akan dihitung titik pusat *cluster* (*centroid*) dari setiap *cluster* yang sudah ditentukan sebelumnya. Dengan menghitung jarak *euclid* setiap objek data dengan titik pusat *cluster* maka akan didapatkan kemiripan karakteristik data untuk mengelompokkan anggota *cluster*. Objek dengan jarak *euclid* yang mendekati *centroid* tertentu akan menggambarkan kemiripan karakteristik dalam *cluster* tersebut sehingga akan didapatkan sekelompok anggota *cluster* yang mirip dengan *centroid*nya.

2.2. *K-Means*

Metode *K-Means* akan mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok memiliki karakteristik yang berbeda dengan data pada kelompok lainnya namun memiliki kemiripan karakteristik dalam satu kelompoknya. Algoritma dasar yang akan dilakukan pada metode ini sebagai berikut:

a. Jumlah *cluster* akan ditentukan terlebih dahulu dengan metode yang sudah ada, contoh dengan menggunakan metode *Elbow* yang akan menampilkan lekukan dalam mendapatkan jumlah *cluster* terbaiknya.
b. Mendapatkan pusat *cluster* (*centroid*) yang didapat dengan cara *random*.
c. Menghitung jarak *Euclid* setiap objek variabel yang diamati ke pusat *cluster* (*centroid*), rumus yang digunakan adalah *Euclidean Distance* berikut:

$$d_{AB} = \sqrt{\sum_{k=1}^{r} (X_{Ak} - X_{Bk})^2}$$

dengan,

d_{AB} = Jarak objek A ke B

r = Jumlah objek yang digunakan

X_{Ak} = Nilai dari objek A ke k

X_{Bk} = Nilai dari objek B ke k

d. Mengelompokkan objek ke dalam *cluster* yang diharapkan berdasarkan jarak terdekat objek dengan *centroid*nya.
e. Melakukan iterasi berikutnya untuk menentukan posisi *centroid* baru.
f. Melakukan perhitungan *Euclidean Distance* kembali jika posisi *centroid* baru dengan *centroid* lama berbeda.

(Harapan, Baginda, 2019).

Algoritma *K-Means* dibangun berdasarkan kemiripan anggota dalam kelompok dengan mengukur jarak *euclid* suatu objek dengan pusat *cluster* yang sudah terbentuk (Fitri, Hidayatul, Asroni, Eko Prasetyo, 2018).

2.3. Data Mining

Data mining merupakan proses dalam menemukan pola atau informasi menarik pada objek data yang dipilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma pada data mining sangat beragam. Pemilihan metode atau algoritma yang benar bergantung pada keseluruhan tujuan dan proses penemuan pengetahuan dalam *database* (proses penggalian informasi tersembunyi dalam database besar). Sebenarnya keduanya memiliki konsep yang berbeda, namun saling berkaitan (SY, Hasyif, Rismayani, Asrul Syam, 2019).

2.4. Metode Penelitian

Bahan penelitian yang digunakan adalah data jumlah tenaga kesehatan di provinsi Banten pada tahun 2020. Data tersebut dilakukan dikelompokkan (*clustering*) dengan menggunakan algoritma *K-Means* dengan bantuan bahasa pemograman *python* untuk mengelompokkan objek data pada *k cluster* yang akan dibentuk. Data tersebut juga menggunakan output *google colab* atau *google colaboratory*. *Google colab* merupakan *tools* yang berbasis *cloud* yang disediakan oleh *google* dengan gratis kepada penggunanya yang bertujuan untuk *research*/penelitian.

3. PEMBAHASAN

Data penelitian menggunakan data sekunder yang diperoleh melalui data jumlah tenaga kesehatan di Provinsi Banten pada tahun 2020 yang dikeluarkan oleh Badan Pusat Statistika (BPS) pada tahun 2021. Variabel yang digunakan dibatasi pada 5 variabel tenaga kesehatan yaitu: dokter, dokter gigi, perawat, bidan, dan tenaga kefarmasian. Untuk variabel tenaga kesehatan lain lebih lengkap dapat dilakukan penelitian lanjutan.

Tabel 1. Data Jumlah Tenaga Kesehatan Provinsi Banten

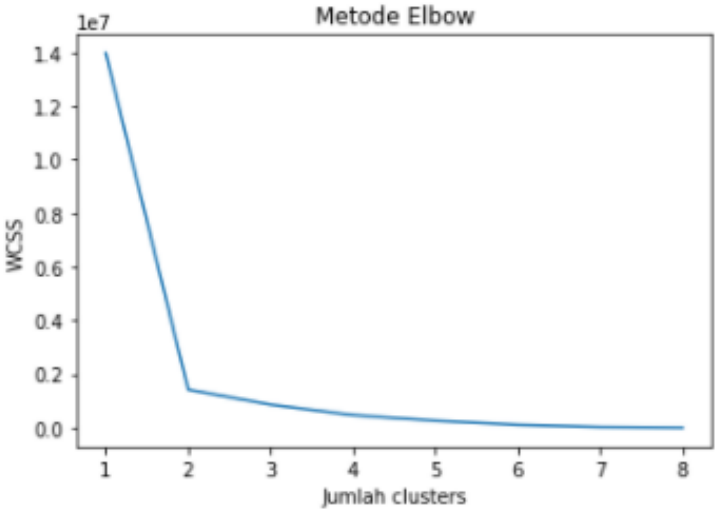
Kabupaten/Kota	Dokter	Dokter Gigi	Perawat	Bidan	Tenaga Kefarmasian
	Kabupaten				
Pandeglang	182	25	698	964	93
Lebak	335	44	1365	1224	153
Tangerang	1670	150	3280	1456	485
Serang	367	42	1099	795	135

Kabupaten/Kota	Dokter	Dokter Gigi	Perawat	Bidan	Tenaga Kefarmasian
Kota					
Tangerang	1860	215	3461	984	570
Cilegon	427	69	926	398	264
Serang	377	48	956	583	159
Tangerang Selatan	1849	224	2536	855	463
Banten	7067	817	14321	7299	2322

(Sumber: Badan Pusat Stastistika Provinsi Banten, 2021)

3.1. Pengolahan Data

Sebelum melakukan iterasi penentuan *centroid* awal dan anggota *cluster*, maka dilakukan penentuan jumlah *cluster* terbaik dengan menggunakan metode Elbow.



Gambar 1. Metode Elbow dengan *Python*

Metode Elbow menunjukkan bahwa terlihat lekukan tanjam pada jumlah *cluster* 2 sehingga jumlah *cluster* terbaiknya adalah 2 *cluster*, maka akan ditentukan *centroid* awal dari 2 *cluster* tersebut. Selanjutnya akan dilakukan proses iterasi untuk menentukan *centroid* awal dan anggota *cluster* awal hingga iterasi yang akan berulang.

Tabel 2. Iterasi 1

No.	Kabupaten/Kota	Dokter	Dokter Gigi	Perawat	Bidan	Tenaga Kefarmasian	C0	C1	Jarak Terdekat	Kelompok Data
1	Kab. Pandeglang	182	25	698	964	93	7899876	3568293	3568293	Cluster 1
2	Kab. Lebak	335	44	1365	1224	153	4655471	1637416	1637416	Cluster 1
3	Kab. Tangerang	1670	150	3280	1456	485	267185	920876	267185	Cluster 0
4	Kab. Serang	367	42	1099	795	135	5835412	2210759	2210759	Cluster 1
5	Kota Tangerang	1860	215	3461	984	570	0	883807	0	Cluster 0
6	Kota Cilegon	427	69	926	398	264	6886006	2865997	2865997	Cluster 1
7	Kota Serang	377	48	956	583	159	6634119	2695248	2695248	Cluster 1
8	Kota Tangerang Selatan	1849	224	2536	855	463	883807	0	0	Cluster 1

Tabel 3. Iterasi 2

No.	Kabupaten/Kota	Dokter	Dokter Gigi	Perawat	Bidan	Tenaga Kefarmasian	C0	C1	Jarak Terdekat	Kelompok Data
1	Kab. Pandeglang	182	25	698	964	93	7422971,8	362373,44	362373,44	Cluster 1
2	Kab. Lebak	335	44	1365	1224	153	4182908,8	192056,44	192056,44	Cluster 1
3	Kab. Tangerang	1670	150	3280	1456	485	66843,75	4574776,1	66843,75	Cluster 0
4	Kab. Serang	367	42	1099	795	135	5515531,8	34207,111	34207,111	Cluster 1
5	Kota Tangerang	1860	215	3461	984	570	66843,75	5011978,1	66843,75	Cluster 0
6	Kota Cilegon	427	69	926	398	264	6734916,8	280947,78	280947,78	Cluster 1
7	Kota Serang	377	48	956	583	159	6390849,8	146608,11	146608,11	Cluster 1
8	Kota Tangerang Selatan	1849	224	2536	855	463	835581,75	1709148,4	835581,75	Cluster 0

Tabel 4. Iterasi 3

No.	Kabupaten/Kota	Dokter	Dokter Gigi	Perawat	Bidan	Tenaga Kefarmasian	C0	C1	Jarak Terdekat	Kelompok Data
1	Kab. Pandeglang	182	25	698	964	93	5952412,7	131082,88	131082,88	Cluster 1
2	Kab. Lebak	335	44	1365	1224	153	3148745	312877,88	312877,88	Cluster 1
3	Kab. Tangerang	1670	150	3280	1456	485	165855	5715521,1	165855	Cluster 0
4	Kab. Serang	367	42	1099	795	135	4228274,7	8848,88	8848,88	Cluster 1
5	Kota Tangerang	1860	215	3461	984	570	153498,67	6247505,7	153498,67	Cluster 0
6	Kota Cilegon	427	69	926	398	264	5259610,7	174010,08	174010,08	Cluster 1
7	Kota Serang	377	48	956	583	159	4973316,3	46852,28	46852,28	Cluster 1
8	Kota Tangerang Selatan	1849	224	2536	855	463	371388,33	2460871,5	371388,33	Cluster 0

Pada iterasi kedua dan ketiga menghasilkan hasil yang sama maka iterasi dihentikan dengan hasil *cluster* dan anggotanya.

3.2. Tahap Awal

Pengelompokkan data yang digunakan *K-Means* mempunyai beberapa tahapan. Tahapan awal proses ini adalah penentuan jumlah *cluster*, penentuan *cluster* yang dibuat atau dihasilkan dalam penelitian ini sebanyak dua *cluster*.

3.3. Penentuan Pusat *Cluster*

Menentukan nilai pusat c*luster* (*centroid*). Dalam menentukan nilai pusat *cluster* penelitian akan menentukan nilai pusat awal yang dilakukan secara random dan didapatkan nilai pusat dari setiap *cluster*.

3.4. Menghitung Jarak Terdekat

Langkah-langkah untuk menghitung jarak antara titik pusat *cluster* dan data yang dimiliki (data tenaga kesehatan) dilakukan dengan menggunakan rumus jarak *Euclidean*. Metode yang dimaksud adalah metode titik terdekat antara dua benda dalam perhitungan jarak, metode ini disebut juga *Euclidean*. Berdasarkan data dari pengolahan data,

didapatkan nilai pusat *cluster* (*centroid*) pada iterasi kedua dan ketiga yang mempunyai nilai yang sama maka proses perhitungan dianggap cukup jika nilai *centroid* tidak berubah. Kasus ini adalah proses iteratif yang diulang tiga kali.

3.5. *Google Colab* atau *Google Colaboratory*

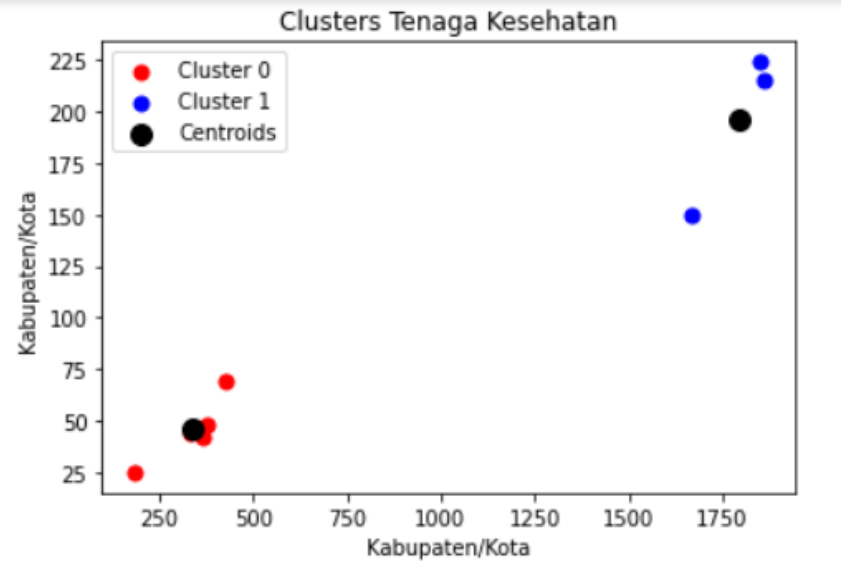
Dalam pengolahan data tersebut juga menggunakan output *google colab* atau *google colaboratory*. *Google colab* merupakan *tools* yang berbasis *cloud* yang disediakan oleh *google* dengan layanan GPU gratis kepada penggunanya yang bertujuan untuk *research*/penelitian yang membutuhkan spesifikasi komputer yang tinggi dalam *running* nya. Dengan menggunakan *google* tersebut juga menyediakan *coding enveronment* bahasa pemrograman *python*, dengan kata lainnya bahwa *google* seperti meminjamkan komputer para peneliti dengan dilengkapi GPU secara *online* dan gratis untuk menghitung atau membuat program dengan sebelumnya melakukan *login* dengan akun *gmail*nya. Untuk versi *offline* nya dapat men*download Jupiter Notebook* untuk menjalankan bahasa pemograman *pyhon* di PC atau laptop.

Hasil dari pengolahan data dengan menggunakan *colab*, bisa dilihat dari gambar 3.4 output dari *cluster* terbentuk yaitu mendapatkan nilai titik *cluster* yang dilakukan pengklasteran sebanyak dua kali.

```
*************** 2 Cluster Model ***************
Cluster centers:
 [[ 337.6          45.6        1008.8          792.8          160.8        ]
 [1793.          196.33333333 3092.33333333 1098.33333333  506.        ]]
Inertia (WCSS):  1421232.7999999998
No. of iterations:  2
```

Gambar 2. Output dari *cluster* terbentuk dengan *Python*

Memasukkan data ke dalam setiap *cluster*, dimana dalam memasukkan data ke dalam setiap *cluster* yang harus dilakukan adalah dengan cara menghitung jarak *euclide* setiap objek data. Langkah selanjutnya adalah mengelompokkan data berdasarkan hasil minimum jarak. Berdasarkan hasil pengolahan data dengan *python*, iterasi dilakukan dua kali hingga diperoleh nilai *centroid* dan mengelompokkan objek data pada setiap *cluster* hingga objek tidak berpindah *cluster*. Untuk hasil *clustering* didapat bisa dilihat dari gambar 3.5 output dari *cluster* yang akan terbentuk dalam sebaran *cluster*.



Gambar 3. Output dari *cluster* yang terbentuk dalam sebaran *cluster*

```
cluster2_label={'CLUSTER':kmeans2.labels_}
cluster2_df=pd.DataFrame(cluster2_label)
cluster2=kabkota1.join(cluster2_df)
print("****** 2 Cluster ***** \n",cluster2)
```

```
****** 2 Cluster *****
        Kabupaten/Kota  CLUSTER
0           Pandeglang        0
1                Lebak        0
2  Kabupaten Tangerang        1
3     Kabupaten Serang        0
4       Kota Tangerang        1
5              Cilegon        0
6          Kota Serang        0
7    Tangerang Selatan        1
```

Gambar 4. Output dari *cluster* yang terbentuk dua *cluster*

Hasil dari data yang dilakukan menentukan bahwa kelompok *cluster* satu hanya tiga yaitu kabupaten Tangerang, kota Tangerang dan kota Tangerang Selatan dan selebihnya masuk pada cluster ke nol. Pusat cluster ke dua merupakan daerah tingkat kemiripan tinggi, cluster yang dihasilkan ada dua *cluster*, yaitu *cluster* 0 (*centroid*: 337.6; 45.6; 1008.8; 792.8; 160.8) dan *cluster* 1 (*centroid*: 1793; 196.3; 3092.3; 1098.3; 506). Hasil yang didapat bisa dilihat dari gambar 3.6 output dari *cluster* yang terbentuk dua *cluster*.

4. SIMPULAN

Cluster dari hasil proses yang dilakukan pada data tenaga kesehatan pada tahun 2020 di provinsi Banten, berdasarkan data tenaga kesehatan peminat lebih cenderung berada di kelompok *cluster* satu. Kelompok *cluster* satu hanya ada tiga bagian yaitu dan selebihnya masuk pada cluster ke nol. Pusat *cluster* ke dua merupakan daerah tingkat kemiripan tinggi berdasarkan lima variabel. *Cluster* yang dihasilkan ada dua *cluster*, yaitu:

1. *Cluster* 0 (*centroid*: 337.6; 45.6; 1008.8; 792.8; 160.8) dengan anggota lima daerah yaitu: kabupaten Pandeglang, kabupaten Lebak, kabupaten Serang, kota Cilegon, dan kota Serang.
2. *Cluster* 1 (*centroid*: 1793; 196.3; 3092.3; 1098.3; 506) beranggota tiga daerah yaitu: kabupaten Tangerang, kota Tangerang dan kota Tangerang Selatan.

5. DAFTAR PUSTAKA

Rahmawati, Rika, Fatayat. (2020). "*Analisa Mahasiswa Mengikuti Badan Eksekutif Mahasiswa Terhadap Prestasi Akademik Menggunakan Metode K-Means Clustering*". Jurnal Sistem Informasi. Pekanbaru.

Fitri, Hidayatul, Asroni, Eko Prasetyo. (2018). " *Penerapan Metodologi Algoritma K-Means pada Pengelompokan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta*". Jurnal Semesta Teknika Vol.21 No.1. Yogyakarta.

SY, Hasyif, Rismayani, Asrul Syam. (2019). "Data Mining Menggunakan Algoritma K-Means Pengelompokan Penyebaran Diare di Kota Makassar". Jurnal Prosiding Seminar Ilmiah Sistem Informasi dan Teknologi Informasi Vol.VIII No.1. Makassar.

Harapan, Baginda. (2019). "*Penerapan Algoritma K-Means untuk Menentukan Bahan Bangunan Laris*". Jurnal Teknologi dan Bisnis. Sumatera Utara.

Badan Pusat Stastistika Provinsi Banten. (2021). "*Provinsi Banten Dalam Angka 2021*". Banten