Pengelompokkan Provinsi di Indonesia Berdasarkan Penyakit Tidak Menular Menggunakan Metode Partisi, Hierarki, dan Fuzzy Clustering

Salsavira
Politeknik Statistika STIS, salsvir25@gmail.com

ABSTRAK, *Penyakit tidak menular menjadi isu penting di Indonesia. Berdasarkan data Riset Kesehatan Dasar, tingkat prevalensi penyakit tidak menular di Indonesia melonjak lebih dari 34 persen pada tahun 2018. Oleh karena itu diperlukan suatu penelitian sebagai dasar pemerintah dan pihak terkait untuk membuat kebijakan yang tepat. Penelitian pengelompokkan provinsi di Indonesia berdasarkan penyakit tidak menular menggunakan metode partisi, hierarki, dan fuzzy clustering memiliki tujuan mengelompokkan provinsi-provinsi di Indonesia berdasarkan penyakit tidak menular dan menentukan metode terbaik untuk melakukan pengelompokkannya. Variabel yang digunakan pada penelitian ini terdiri dari delapan jenis penyakit tidak menular yang bersumber dari data RISKESDAS 2018. Berdasarkan hasil pengujiannya didapatkan bahwa provinsi-provinsi di Indonesia dapat dikelompokkan berdasarkan penyakit tidak menular dengan menggunakan metode k-means, k-medoids, AGNES, DIANA, dan fuzzy c-means clustering, dan berdasarkan hasil perbandingan kelima metode tersebut dapat disimpulkann bahwa metode paling baik untuk mengelompokkan provinsi di Indonesia berdasarkan penyakit tidak menular adalah dengan fuzzy c-means clustering.*

***Kata Kunci**: Penyakit Tidak Menular, K-Means, K-Medoids, AGNES, DIANA, Fuzzy C-Means Clustering*

1. PENDAHULUAN

Kesehatan merupakan suatu hal yang penting dalam kelangsungan hidup manusia. Tanpa kesehatan manusia akan sulit bahkan tidak bisa melakukan kegiatan sehari-hari. Namun tentunya manusia pasti tidak akan pernah terhindar dari penyakit. Penyakit sendiri dapat diklasifikasikan menjadi berbagai macam dan berdasarkan sifat penularannya penyakit dapat dibagi menjadi penyakit menular dan penyakit tidak menular. Darmawan [5] dalam jurnalnya mendefinisikan penyakit menular sebagai penyakit yang terjadi akibat adanya interaksi antara agen penyakit berupa mikroorganisme hidup, manusia, dan lingkungan, sedangkan penyakit tidak menular didefinisikan sebagai penyakit yang terjadi akibat interaksi antara agen penyakit berupa *non living agent*, manusia dan lingkungan.

Seringkali penyakit menular dianggap lebih berbahaya dibandingkan penyakit tidak menular karena resiko seseorang untuk terkena penyakit menular lebih tinggi, padahal berdasarkan fakta yang didapatkan dari Badan Kesehatan Dunia (WHO) [8] penyakit tidak menular telah membunuh 41 juta orang tiap tahun atau setara dengan 71 persen dari jumlah kematian global. Setiap tahunnya 15 juta orang yang berusia 30 sampai 69 tahun meninggal dikarenakan penyakit tidak menular dan kejadian kematian tersebut paling banyak terjadi di negara berpenghasilan rendah dan menengah.

Penyakit tidak menular juga menjadi isu penting di Indonesia. Hal ini dikarenakan penderita penyakit tidak menular di Indonesia terus meningkat setiap tahunnya. Berdasarkan data Riset Kesehatan Dasar, tingkat prevalensi penyakit tidak menular melonjak lebih dari 34 persen di Indonesia. Peningkatan terbesar terjadi pada penyakit hipertensi dimana pada tahun 2018 terjadi kenaikan sebesar 8,3 persen. Senada dengan hipertensi, penderita kanker, diabetes melitus, penyakit jantung, stroke, dan gagal ginjal kronis juga mengalami peningkatan pada tahun 2018. Di lain sisi, penyakit asma dan penyakit sendi mengalami penurunan apabila dibandingkan dengan data pada tahun 2013. Pada tahun 2013, prevalensi asma berdasarkan diagnosis dokter sebesar 4,5 persen dan menurun menjadi 2,4 persen pada 2018. Sedangkan pada penyakit sendi terjadi penurunan sebesar 4,6 persen pada tahun 2018.

Machine learning merupakan suatu kecerdasan buatan berupa penerapan algoritma yang bertujuan untuk melakukan prediksi, pengenalan pola, dan klasifikasi [2]. Pada bidang kesehatan, penggunaan analisis dengan *machine learning* telah banyak dilakukan untuk pengembangan obat-obatan serta pengembangan teknologi untuk memudahkan diagnosis penyakit pada pasien.

Pentingnya mengelompokkan provinsi di Indonesia berdasarkan penyakit tidak menular menjadi hal yang penting untuk dilakukan agar

pemerintah dan pihak terkait dapat membuat kebijakan yang tepat dan memperketat pengawasan untuk mengatasi peningkatan penyakit tidak menular di Indonesia, terutama pada daerah yang memiliki tingkat penyakit tidak menular yang tinggi. Pada penelitian ini dilakukan pengelompokkan provinsi di Indonesia berdasarkan penyakit tidak menular dengan menggunakan metode partisi, hierarki, dan *fuzzy clustering* untuk mengelompokkan provinsi-provinsi di Indonesia berdasarkan penyakit tidak menular dan menentukan metode terbaik untuk melakukan pengelompokkannya. Hasil dari penelitian ini diharapkan dapat digunakan sebagai dasar untuk penelitian-penelitian selanjutnya, sehingga *machine learning* dapat terus berkembang dalam dunia kesehatan terutama untuk mengatasi peningkatan penyakit tidak menular di Indonesia.

2. TINJAUAN PUSTAKA

ANALISIS KLASTER

Analisis klaster adalah salah satu analisis multivariat. Analisis klaster merupakan *unsupervised learning* karena merupakan proses pembelajaran tanpa adanya label dari objek. Analisis klaster bertujuan untuk mengelompokkan objek-objek dengan karakteristik yang mirip menjadi satu kelompok dan objek dengan karakteristik yang berbeda menjadi kelompok yang lain [9]. Analisis klaster melakukan proses *learning* berdasarkan karakteristik atau pola natural dari data. Sebuah klaster yang baik memiliki ciri sebagai berikut [1]:

a. Homogenitas (kemiripan) yang tinggi antar objek dalam satu klaster (*within-cluster*)
b. Heterogenitas (perbedaan) yang tinggi antara klaster yang satu dengan yang lainnya (*between cluster*)

METODE PENGELOMPOKKAN

Berdasarkan nilai keanggotaan yang terbentuk, tipe pengelompokkan dalam analisis klaster terbagi menjadi dua yakni [4]:

a. *Soft Clustering* (*Overlapping Clustering*)
   Pada tipe pengelompokkan *soft clustering*, tiap titik atau objek dapat memiliki dua atau lebih klaster dengan derajat keanggotaan yang berbeda. Tiap objek tidak dipaksa sepenuhnya menjadi satu kelompok, melainkan keanggotaan suatu objek dalam suatu klaster direpresentasikan dengan nilai peluang antara 0 dan 1. Penentuan anggota klaster ditentukan dengan memilih peluang terbesar.
b. *Hard Clustering* (*Exclusive Clustering*)
   Pada tipe pengelompokkan *hard clustering*, sebuah objek tidak diperbolehkan untuk menjadi anggota dari dua kelompok atau lebih, sehingga dapat dipastikan bahwa suatu objek hanya menjadi anggota satu klaster tertentu.

Beberapa metode atau pendekatan yang digunakan dalam tipe pengelompokkan *hard clutering* antara lain:

a. Pendekatan partisi, pendekatan pengelompokkan objek dengan memaksimalkan ukuran kriteria yang akan digunakan sehingga *Sum Squared Distance*-nya akan mengecil.
b. Pendekatan hierarki, pendekatan pengelompokkan dengan membuat suatu hierarki berupa dendogram.
c. Pendekatan kepadatan, pendekatan klaster dengan menemukan klaster yang merupakan wilayah yang padat pada data dan dipisahkan oleh data atau objek yang renggang.
d. Pendekatan *grid*, pendekatan klaster dengan menggunakan *grid* sebagai objek pengamatannya.

ANALISIS KLASTER PARTISI

Analisis klaster partisi akan mempartisi atau membagi data menjadi beberapa klaster sehingga *Sum of Square* terminimalkan. Pendekatan partisi dibagi menjadi dua metode yakni:

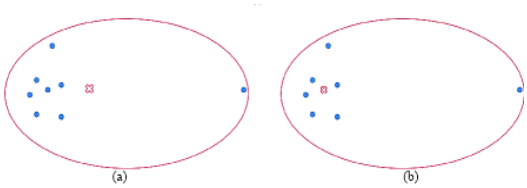
a. *Global optimal*, metode ini menghitung satu demi satu *all possible combination*.
b. Metode heuristik, metode partisi yang dilakukan dengan *trial* dan *error*.

Metode heuristik merupakan metode dalam analisis klater partisi yang sering digunakan. Adapun metode-metode yang terdapat di dalam metode heuristik antara lain:

a. *K-means*, merupakan metode yang masing-masing klasternya

direpresentasikan oleh pusat klaster. Pada *k-means*, pusat klaster dapat berupa ruang kosong maupun salah satu objek.

b. *K-medoids*, merupakan metode yang mirip dengan *k-means* dimana tiap klaster diwakilkan oleh pusat klasternya. Perbedaan antara *k-medoids* dan *k-means* terletak pada bentuk pusat klasternya, dimana pada *k-medoids* pusat klasternya harus berupa objek atau salah satu data di dalam klaster.



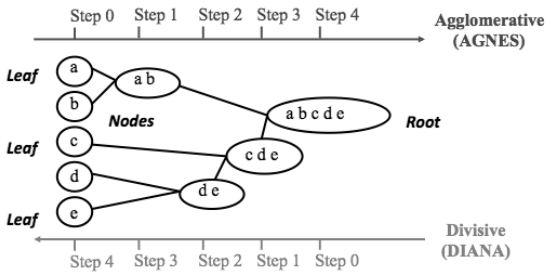
Gambar 2.1 (a) *K-Means* dan (b) *K-Medoids*

ANALISIS KLASTER HIERARKI

Analisis klaster hierarki menempatkan objek yang mirip pada hierarki yang berdekatan, sedangkan objek yang tidak mirip akan ditempatkan pada hierarki yang berjauhan [9]. Pendekatan hierarki dibagi menjadi dua yakni:

a. *Agglomerative hierarchial clustering* (AGNES), metode yang menggunakan strategi *bottom-up* dimana pada awalnya setiap objek merupakan klaster tersendiri dan kemudian objek tersebut digabung menjadi klaster yang lebih besar.

b. *Divisive hierarchial clustering* (DIANA), merupakan metode yang berkebalikan dengan AGNES yakni menggunakan strategi *top-down* dimana seluruh objek pada awalnya merupakan satu klaster besar dan kemudian dipisahkan menjadi klaster yang lebih kecil.



Gambar 2.2 AGNES dan DIANA

ANALISIS KLASTER FUZZY

Analisis klaster *fuzzy* merupakan alat untuk memecahkan masalah ketidakpastian pada *clustering* [6]. Salah satu metode klaster *fuzzy* yang paling banyak digunakan adalah *Fuzzy C-Means*. *Fuzzy C-Means* adalah teknik pengelompokkan data dimana keberadaan titik dalam suatu kelompok ditentukan oleh derajat keanggotannya.

ASUMSI

Kaiser-Meyer Olkin merupakan salah satu asumsi yang harus dipenuhi dalam analisis klaster. Dengan KMO dapat diketahui apakah jumlah data yang ada sudah dapat mewakili populasi.

Hipotesis:
H_0: Sampel belum cukup untuk dilakukan analisis
H_1: Sampel sudah cukup untuk dilakukan analisis

Statistik Uji:

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^{\,2}}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^{\,2} + \sum_{i=1}^{p}\sum_{j=1}^{p} \alpha_{ij}^{\,2}} \qquad (2.1)$$

dimana
r_{ij} = koefisien korelasi sederhana antara variabel i dan j
α_{ij} = koefisien korelasi parsial antara variabel i dan j

Keputusan:
Tolak H_0 jika KMO > 0,5

Selain itu asumsi yang harus dipenuhi adalah asumsi multikolinearitas. Asumsi multikolinearitas merupakan salah satu syarat untuk melakukan analisis multivariat, dimana antar variabel atau atribut tidak boleh terdapat korelasi yang kuat. Multikolinearitas dapat dideteksi dengan menghitung koefisien korelasi Pearson antar variabelnya . Adapun rumus dari korelasi Pearson adalah sebagai berikut.

$$r_{xy} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} \qquad (2.2)$$

dimana
r = koefisien korelasi Peason
n = jumlah sampel

Apabila setelah perhitungan terdapat korelasi yang kuat antara variabel (nilai korelasi melebihi 0,5) maka terindikasi terjadi multikolinearitas. Salah satu metode yang dapat digunakan untuk menyelesaikan masalah multikolinearitas adalah dengan menggunakan *Principal Component Analysis* (PCA).

UKURAN KEMIRIPAN OBJEK

Dalam mengukur kemiripan atau kesamaan antar objek digunakan ukuran jarak (*distance*) sebagai pendekatan. Semakin jauh atau semakin besar jarak antar objek maka semakin besar pula perbedaan karakteristik antar objek tersebut, sedangkan semakin dekat atau semakin kecil jarak antar objek maka semakin mirip karakteristik antar objek tersebut. Ada beberapa jarak yang biasa digunakan antara lain [7]:

a. *Euclidean distance*, merupakan metode penghitungan jarak untuk mengukur jarak dari dua buah titik dalam *Euclidean space*. Rumus yang digunakan untuk menghitung *Euclidean distance* adalah sebagai berikut.

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2.3)$$

dimana
x_i = koordinat x untuk variabel i
y_i = koordinay y untuk variabel i

b. *Manhattan distance*, merupakan metode penghitungan jarak untuk melihat perbedaan mutlak dari koordinat antar dua obek. Rumus yang digunakan adalah sebagai berikut.

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

c. *Minkowski distance*, merupakan sebuah metrik yang dianggap sebagai generalisasi dari *Euclidean distance* dan *Manhattan distance*. Rumus yang digunakan dalam menghitung *Minkowski distance* adalah sebagai berikut

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (2.4)$$

dimana
p = banyak variabel yang diamati

VALIDASI KLASTER

Validasi klaster merupakan suatu usaha yang dilakukan untuk mengevaluasi hasil klaster. Proses validasi memberikan informasi tentang akurasi jumlah klaster yang dipilih. Apabila solusi klaster yang dihasilkan tidak berbeda jauh berdasarkan metode yang digunakan maka dapat dikatakan jumlah klaster yang terbentuk sudah baik. Terdapat tiga pendekatan utama dalam melakukan validasi klaster yaitu [11]:

a. *External test*, hasil *clustering* data *input* dibandingkan dengan hasil *clustering* data yang bukan merupakan data *input*.
b. *Internal test*, hanya data *input* yang digunakan untuk evaluasi kualitas *clustering*. Biasanya digunakan pada pendekatan partisi dan hierarki.
c. *Relative test*, beberapa *clustering* yang berbeda dari satu kumpulan data dibandingkan menggunakan algoritma *clustering* yang sama dengan parameter yang berbeda.

3. METODOLOGI

SUMBER DATA

Data yang digunakan merupakan data 34 provinsi di Indonesia yang bersumber dari Riset Kesehatan Dasar (RISKESDAS) Tahun 2018 yang diterbitkan oleh Kementerian Kesehatan Republik Indonesia. Data yang ada di dalam publikasi ini bersumber dari survei yang dilakukan oleh Kementerian Kesehatan dan bekerjasama dengan Badan Pusat Statistik.

VARIABEL PENELITIAN

Dalam penelitian ini digunakan delapan variabel penelitian yang merupakan indikator penyakit tidak menular di Indonesia. Penjelasan lebih lanjut mengenai variabel yang akan digunakan dapat dilihat pada tabel 1.

Tabel 3.1 Variabel Penelitian

No	Variabel	Keterangan
1	Asma	Prevalensi asma berdasarkan diagnosis dokter pada penduduk semua umur

2	Kanker	Prevalensi (per mil) kanker berdasarkan diagnosis dokter pada penduduk semua umur
3	Diabetes melitus	Prevalensi diabetes melitus berdasarkan diagnosis dokter pada penduduk semua umur
4	Penyakit jantung	Prevalensi penyakit jantung berdasarkan diagnosis dokter pada penduduk semua umur
5	Hipertensi	Prevalensi hipertensi berdasarkan diagnosis dokter atau minum obat antihipertensi pada penduduk umur ≥ 18 tahun
6	Stroke	Prevalensi (per mil) stroke berdasarkan diagnosis dokter pada penduduk umur ≥ 15 tahun
7	Gagal ginjal kronis	Prevalensi gagal ginjal kronis berdasarkan diagnosis dokter pada penduduk umur ≥ 15 tahun
8	Penyakit sendi	Prevalensi penyakit sendi berdasarkan diagnosis dokter pada penduduk umur ≥ 15 tahun

4. PEMBAHASAN

STATISTIK DESKRIPTIF

Gambaran mengenai variabel penyakit tidak menular yang digunakan dapat dilihat dalam tabel 4.1.

Tabel 4.1 Statistik Deskriptif Variabel Penyakit Tidak Menular di Indonesia

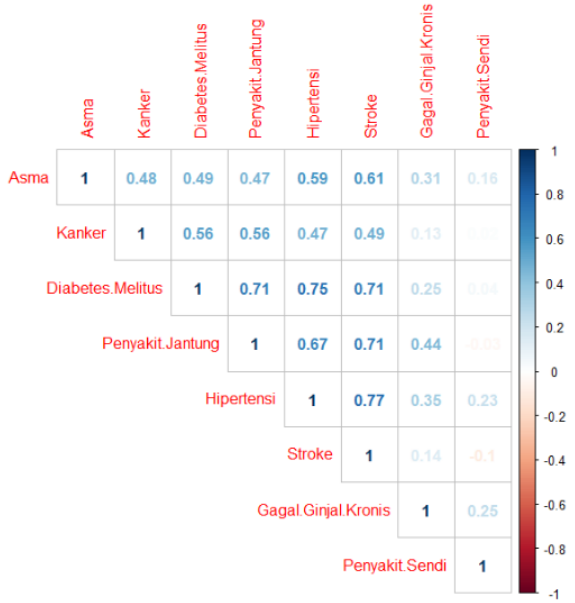
Variabel	Min	Max	Rata-rata
Asma	1,00	4,50	2,485
Kanker	0,85	4,86	1,759
Diabetes melitus	0,60	2,60	1,376
Penyakit jantung	0,70	2,20	1,438
Hipertensi	4,39	13,21	8,181
Stroke	4,10	14,70	10,080
Gagal ginjal kronis	0,18	0,64	0,395
Penyakit sendi	3,16	13,26	7,119

Berdasarkan tabel di atas dapat diketahui nilai minimum, maksimum, serta rata-rata setiap variabel. Penyakit tidak menular yang banyak terjadi di Indonesia pada tahun 2018 adalah stroke dengan rata-rata 10,08 persen. Lalu disusul oleh hipertensi dan penyakit sendi dimana kedua jenis penyakit tidak menular tersebut memiliki rata-rata lebih dari lima persen di Indonesia. Selain itu, dari tabel dapat pula dilihat bahwa penyakit tidak menular yang paling jarang terjadi adalah penyakit gagal ginjal kronis dimana rata-ratanya hanya mencapai 0,395 persen pada tahun 2018.

UJI ASUMSI

Hasil uji KMO menunjukkan bahwa nilai KMO sebesar 0,77. Nilai KMO > 0,5. Hal ini menyebabkan keputusan tolak H_0 dan dapat disimpulkan bahwa sampel yang digunakan sudah cukup sehingga dapat dilakukan analisis lebih lanjut.

Pendeteksian multikolinearitas dilakukan dengan melihat korelasi antar variabel. Adapun korelasi antar variabel yang digunakan dapat dilihat pada gambar berikut.



Gambar 4.1 Korelasi antar Variabel

Dari plot dapat diketahui bahwa terdapat multikolinearitas dikarenakan terdapat nilai korelasi ≥ 0,5 yang mengindikasikan adanya korelasi yang kuat antar variabel. Contohnya saja pada variabel hipertensi dan stroke, penyakit jantung dan stroke, dan diabetes melitus dan hipertensi. Oleh karena itu, untuk mengatasi adanya multikolinearitas dilakukan analisis komponen utama atau *Principal Component Analysis* (PCA).

Asumsi yang harus terpenuhi untuk melakukan PCA adalah sebagai berikut:

a. Uji Kecukupan Sampel
   Kesimpulan dari nilai KMO yang didapatkan adalah sampel yang

digunakan sudah cukup untuk dilakukan analisis lebih lanjut.

b. Uji Bartlett

Hasil dari uji Bartlett dapat dilihat pada tabel 3 berikut ini.

Tabel 4.2 Uji Bartlett

Uji		Nilai
Bartlett's Test of Sphericity	*Approx. Chi-Square*	135,11
	Df	28
	Sig	5,537e-16

Berdasarkan tabel dapat dilihat bahwa p-value sebesar 5,537e-16. Nilai p-value < 0,05 sehingga dapat diambil keputusan tolak H_0 dan dapat disimpulkan bahwa terdapat korelasi antar variabel.

c. Nilai MSA

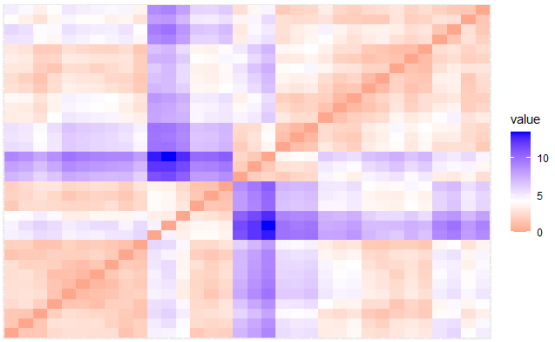
Tabel 4.3 Nilai MSA

Variabel	MSA
Asma	0,81
Kanker	0,83
Diabetes melitus	0,89
Penyakit jantung	0,79
Hipertensi	0,80
Stroke	0,74
Gagal ginjal kronis	0,52
Penyakit sendi	0,32

Berdasarkan tabel dapat dilihat bahwa variabel penyakit sendi memiliki nilai MSA < 0,5. Sehingga variabel tersebut dapat dihapus.

PREDIAGNOSTIC

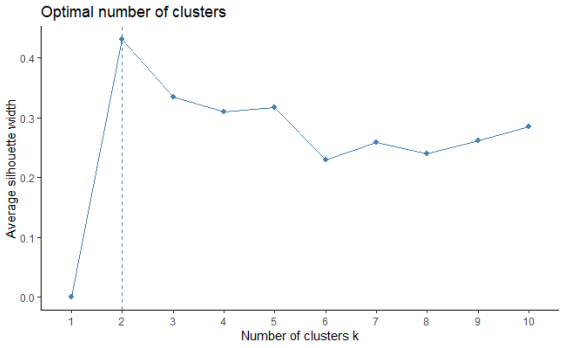
Setelah dilakukan pengujian dengan Statistik Hopkins didapatkan nilai sebesar 0,68. Nilai yang didapatkan bermakna bahwa data bersifat acak dan dapat dibagi menjadi beberapa klaster.



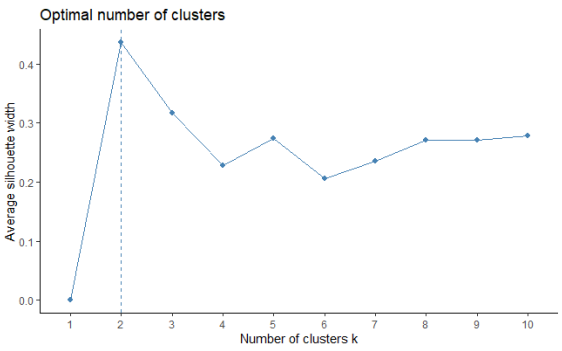
Gambar 4.2 Visualisasi Sebaran Data

JUMLAH KLASTER OPTIMUM

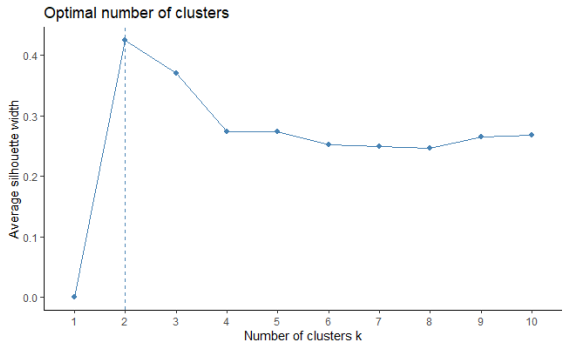
Pemilihan jumlah klaster optimum dilakukan dengan pendekatan *average silhouette width*. Pada penelitian ini, *average silhouette width* menunjukkan klaster optimum untuk seluruh metode yang digunakan adalah sama yakni sebanyak 2. Hal ini dikarenakan pada clustering dengan k=2 menghasilkan nilai *average silhouette width* mendekati 1 yang artinya data akan terkelompokkan dengan baik.



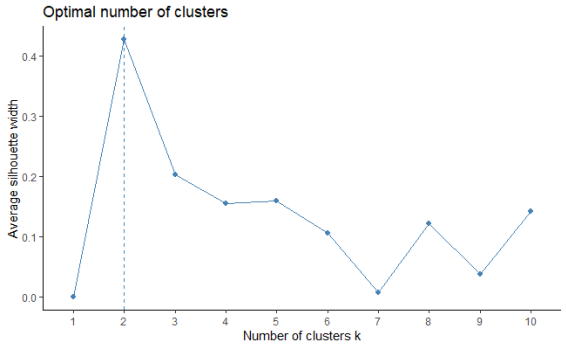
(a)



(b)



(c)

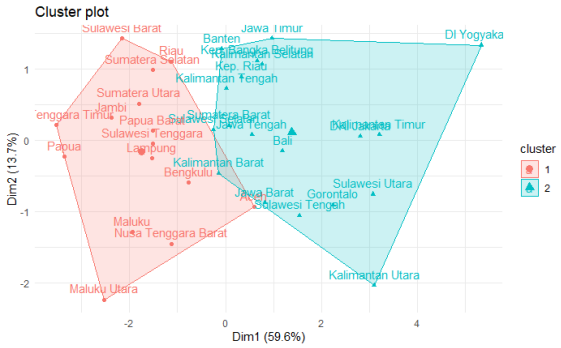


(d)

Gambar 4.3 Klaster Optimum dengan *Average Silhouette Width* pada (a)*K-Means*, (b)*K-Medoids*, (c)Klaster Hierarki, dan (d)*Fuzzy Clustering*

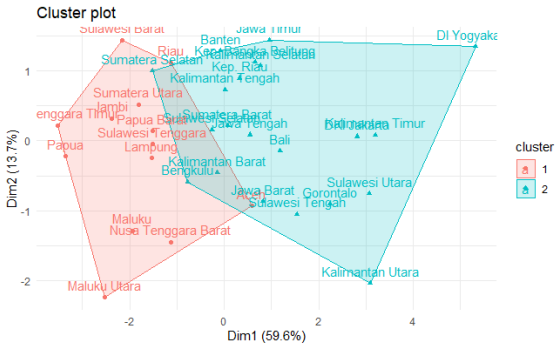
ANALISIS KLASTER PARTISI

Hasil analisis klaster metode *k-means* dengan menggunakan jarak Euclidean membagi provinsi-provinsi di Indonesia menjadi dua klaster, klaster 1 terdiri dari 15 provinsi dan klaster 2 terdiri dari 19 provinsi. Hasil analisis ditampilkan pada gambar 4.4.



Gambar 4.4 *Clusplot K-Means*

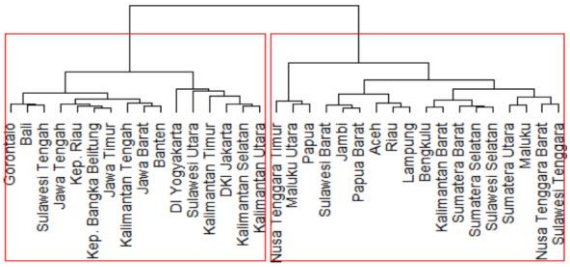
Hasil analisis klaster metode *k-medoids* dengan menggunakan jarak Euclidean membagi provinsi-provinsi di Indonesia menjadi dua klaster, klaster 1 terdiri dari 13 provinsi dan klaster 2 terdiri dari 21 provinsi. Hasil analisis ditampilkan pada gambar 4.5



Gambar 4.5 *Clusplot K-Medoids*

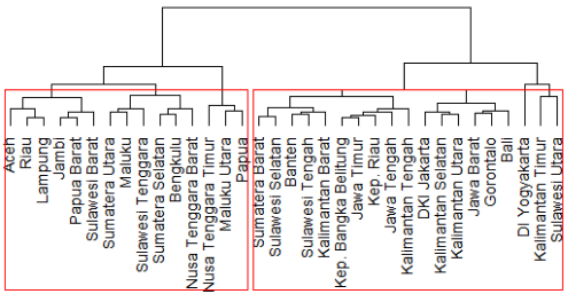
ANALISIS KLASTER HIERARKI

Hasil analisis *agglomerative hierarchial clustering* (AGNES) dengan jarak Euclidean dan metode Ward membagi provinsi di Indonesia menjadi dua klaster, klaster 1 terdiri dari 18 provinsi dan klaster 2 terdiri dari 16 provinsi. Hasil pengelompokkan yang terbentuk ditampilkan melalui dendogram berikut.



Gambar 4.6 Dendogram *Agglomerative Hierarchial Clustering* (AGNES)

Hasil analisis *divisive hierarchial clustering* (DIANA) dengan jarak Euclidean dan metode Ward membagi provinsi di Indonesia menjadi dua klaster, klaster 1 terdiri dari 15 provinsi dan klaster 2 terdiri dari 19 provinsi. Hasil pengelompokkan yang terbentuk ditampilkan melalui dendogram berikut.



Gambar 4.7 Dendogram *Divisive Hierarchial Clustering* (DIANA)

ANALISIS KLASTER FUZZY

Pada tahap penentuan jumlah klaster optimum telah diketahui bahwa pada *fuzzy c-*

means clustering, jumlah klaster yang paling baik untuk digunakan adalah sebanyak dua klaster. Setelah jumlah klaster ditentukan maka tahapan selanjutnya adalah menentukan parameter *fuzzy*. Bezdek et. al. [9] merekomendasikan nilai m yang berkisar antara 1,5 hingga 3. Perbandingan yang dilakukan untuk memilih parameter *fuzzy* terbaik dilakukan dengan kriteria validasi internal berikut:

d. Koefisien Dunn, merupakan koefisien yang dihitung dengan mengukur rasio jarak terbesar antar klaster dengan jarak terkecil di dalam klaster [3]. Semakin tinggi koefisien Dunn maka semakin baik klaster terbentuk.
e. *Within Cluster Sum of Square*, merupakan ukuran yang menunjukkan varians klaster. Semakin kecil nilainya maka klaster yang terbentuk semakin baik [10].

Adapun perbandingan hasil antar parameter *fuzzy* pada data penyakit tidak menular dapat dilihat pada tabel di bawah ini.

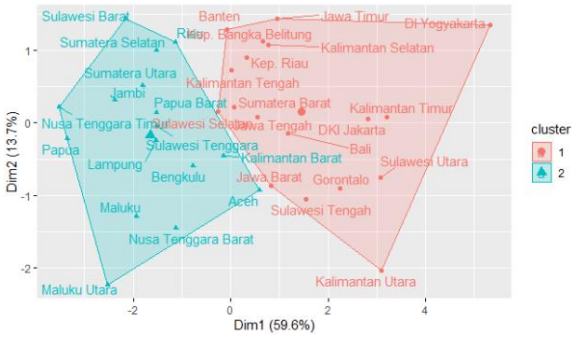
Tabel 4.4 Perbandingan Parameter *Fuzzy* (*Fuzzifier*)

Fuzzifier (m)	Koefisien Dunn	*Within Cluster Sum of Square*
1,2	0,96	56,49%
1,5	0,90	56,87%
2,0	0,77	56,01%
2,5	0,67	54,46%

Pada tabel 9 dapat dilihat bahwa parameter *fuzzy* dengan koefisien Dunn terbesar ketika m=1,2. Sedangkan untuk *Within Cluster Sum of Square*, dapat terlihat bahwa nilai terkecil saat m=2,5.

Berdasarkan hasil perbandingan tersebut dapat disimpulkan bahwa parameter *fuzzy* yang baik untuk digunakan adalah m=1,2. Hal ini dikarenakan pada saat parameter *fuzzy* sebesar 1,2, nilai koefisien Dunn-nya merupakan yang terbesar dan nilai *Within Cluster Sum of Square* yang didapatkan tidak terlalu tinggi.

Hasil analisis *fuzzy c-means clustering* dengan klaster sebanyak dua dan parameter *fuzzy* sebesar 1,2 membagi provinsi di Indonesia menjadi dua klaster, klaster 1 terdiri dari 18 provinsi dan klaster 2 terdiri dari 16 provinsi. Klaster yang terbentuk dapat dilihat pada *plot* berikut.



Gambar 4.8 *Clusplot Fuzzy C-Means*

PERBANDINGAN ANTAR METODE *CLUSTERING*

Penentuan metode terbaik untuk mengelompokkan penyakit tidak menular di Indonesia dilakukan dengan membandingkan hasil *clustering* berbagai metode yang telah dijelaskan sebelumnya. Adapun dalam membandingkan metode-metode tersebut digunakan beberapa kriteria validasi internal yakni:

- Koefisien Dunn
- Koefisien *silhouette*, merupakan koefisien yang mengukur seberapa baik observasi dikelompokkan dengan memperkirakan jarak rata-rata antar klaster. Nilai *silhouette* berkisar antara 1 hingga -1, semakin mendekati 1 maka *clustering* semakin sesuai sedangkan semakin mendekati nilai -1 maka penempatannya semakin tidak sesuai.

Perbandingan antara metode partisi, hierarki, dan *fuzzy clustering* dapat dilihat pada tabel berikut ini.

Tabel 4.5 Perbandingan antar Metode

Metode	Koefisien Dunn	Koefisien *Sillhouette*
K-Means	0,1296	0,4311
K-Medoids	0,1819	0,4368
AGNES	0,1749	0,4250
DIANA	0,1296	0,4311
Fuzzy C-Means	0,9608	0,6683

Dari tabel perbandingan antar metode di atas dapat dilihat bahwa koefisien Dunn terbesar saat menggunakan *fuzzy c-means clustering*. Hal yang sama juga terjadi pada koefisien *silhouette*, dimana nilai yang terbesar atau lebih mendekati satu dihasilkan pada saat menggunakan *fuzzy c-means clustering*. Selain itu didapatkan pula bahwa nilai *Partition Entropy Index* (PEI) pada *fuzzy c-means clustering* adalah sebesar 0,0683 dan nilai *Partition Coefficient Index* (PCI)-nya adalah sebesar 0,9608. Klaster optimal didapatkan jika nilai PEI kecil (mendekati 0) dan nilai PCI besar (mendekati 1) (Mashfuufah & Istiawan, 2018). Oleh karena itu dapat disimpulkan bahwa pada pengelompokkan penyakit tidak menular, metode *clustering* yang paling baik untuk digunakan adalah dengan metode *fuzzy c-means*.

CLUSTER PROFILING

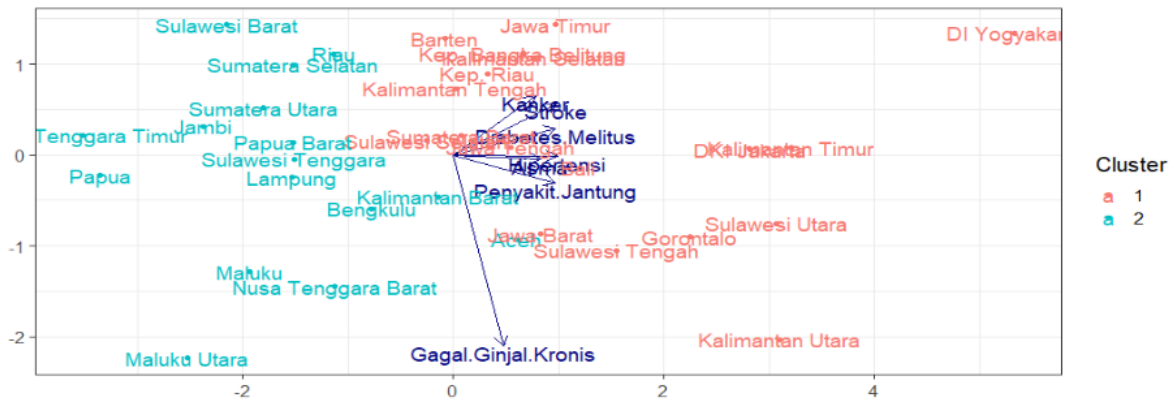
Cluster profiling pada penelitian ini dilakukan dengan menggunakan biplot. Biplot menampilkan gabungan antara observasi yang digambarkan dengan poin dan variabel yang digambarkan dengan panah. Arah panah pada biplot menunjukkan besarnya nilai dari variabel tersebut pada observasi sekitarnya, sedangkan panjang dari panah menggambarkan besarnya varians dari variabel tersebut [9]. Hasil *cluster profiling* menggunakan biplot pada data penyakit tidak menular menggunakan *fuzzy c-means clustering* dapat dilihat pada gambar 4.9.

Biplot tersebut menunjukkan bahwa klaster 1 cenderung memiliki tingkat penyakit tidak menular yang tinggi, sedangkan klaster 2 memiliki tingkat penyakit tidak menular yang rendah. Provinsi-provinsi yang berada di klaster 1 seluruhnya berada di sebelah kanan panah, dimana hal ini mengartikan bahwa tingkat penyakit tidak menular di provinsi-provinsi tersebut nilainya besar. Berbeda dengan provinsi-provinsi di klaster 2 dimana seluruhnya berada di sebelah kiri arah panah, hal ini mengindikasikan bahwa tingkat penyakit tidak menular di provinsi-provinsi tersebut bernilai kecil. Berdasarkan hasil biplot pada gambar 16 dapat disimpulkan bahwa klaster 1 merupakan kelompok provinsi dengan tingkat penyakit tidak menular yang tinggi, sedangkan klaster 2 merupakan kelompok provinsi dengan tingkat penyakit tidak menular yang rendah.

5. KESIMPULAN

Hasil penelitian yang telah dipaparkan memberikan kesimpulan bahwa:

a. Metode partisi *k-means* mengelompokkan provinsi-provinsi di Indonesia berdasarkan penyakit tidak menular menjadi dua klaster, klaster 1 terdiri dari 15 provinsi dan klaster 2 terdiri dari 19 provinsi.
b. Metode partisi *k-medoids* mengelompokkan provinsi-provinsi di Indonesia berdasarkan penyakit tidak menular menjadi dua klaster, klaster 1 terdiri dari 13 provinsi dan klaster 2 terdiri dari 21 provinsi.
c. Metode hierarki *AGNES* mengelompokkan provinsi-provinsi di Indonesia berdasarkan penyakit tidak menular menjadi dua klaster, klaster 1 terdiri dari 18 provinsi dan klaster 2 terdiri dari 16 provinsi.
d. Metode hierarki *DIANA*



Gambar 4.9 Biplot *Fuzzy C-Means Clustering*

mengelompokkan provinsi-provinsi di Indonesia berdasarkan penyakit tidak menular menjadi dua klaster, klaster 1 terdiri dari 15 provinsi dan klaster 2 terdiri dari 19 provinsi.

e. Metode *fuzzy c-means clustering* mengelompokkan provinsi-provinsi di Indonesia berdasarkan penyakit tidak menular menjadi dua klaster, klaster 1 terdiri dari 18 provinsi dan klaster 2 terdiri dari 16 provinsi.
f. Metode terbaik yang dapat digunakan untuk mengelompokkan penyakit tidak menular di Indonesia adalah dengan menggunakan metode *fuzzy c-means clustering*.

6. DAFTAR PUSTAKA

[1] Alwi, W., & Hasrul, M. (2018). Analisis Klaster untuk Pengelompokkan Kabupaten/Kota di Provinsi Sulawesi Selatan berdasarkan Indikator Kesejahteraan Rakyat. Jurnal MSA (Matematika dan Statistika serta Aplikasinya), 6(1), 35.

[2] Bazazeh, D., & Shubair, R. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA) (pp. 1-4). IEEE.

[3] Bhatia, S.K., 2012. A Propound Method For The Improvement of Cluster Quality. IJCSI International Journal of Computer Science Issues, 9(2), 216-221.

[4] Bora, D. J., Gupta, D., & Kumar, A. (2014). A comparative study between fuzzy clustering algorithm and hard clustering algorithm. arXiv preprint arXiv:1404.6059.

[5] Darmawan, A., & Epid, M. (2016). Epidemiologi penyakit menular dan penyakit tidak menular. JAMBI MEDICAL JOURNAL "Jurnal Kedokteran dan Kesehatan", 4(2).

[6] Han J. and Micheline. 2006. Data Mining Consept and Techniques. Morgan Kaufmann. Publishers.

[7] Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. Jurnal Informatika, 4(1).

[8] Non communicable diseases. (2018). Diakses pada 3 December 2020, dari https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[9] Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., & Nooraeni, R. (2018). Data Mining dengan R. IN MEDIA.

[10] Savitri, A. D., Bachtiar, F. A., & Setiawan, N. Y. (2009). Segmentasi Pelanggan Menggunakan Metode K-Means Clustering Berdasarkan Model RFM Pada Klinik Kecantikan (Studi Kasus: Belle Crown Malang). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.

[11] Yatskiv, I., & Gusarova, L. (2005). The methods of cluster analysis results validation. International Conference RelStat, 4, 75-80.