EVALUASI KINERJA METODE CLUSTER ENSEMBLE DAN LATENT CLASS CLUSTERING PADA PEUBAH CAMPURAN*

Debora Chrisinta[1], I Made Sumertajaya[2‡], and Indahwati[3]

[1]Department of Statistics, IPB University, Indonesia, deborachrisinta@gmail.com
[2]Department of Statistics, IPB University, Indonesia, imsjaya.stk@gmail.com
[3]Department of Statistics, IPB University, Indonesia, indah.stk@gmail.com
‡corresponding author

Indonesian Journal of Statistics and Its Applications (eISSN:2599-0802)
Vol 4 No 3 (2020), 448 - 461

Abstract

Most of the traditional clustering algorithms are designed to focus either on numeric data or on categorical data. The collected data in the real-world often contain both numeric and categorical attributes. It is difficult for applying traditional clustering algorithms directly to these kinds of data. So, the paper aims to show the best method based on the cluster ensemble and latent class clustering approach for mixed data. Cluster ensemble is a method to combine different clustering results from two sub-datasets: the categorical and numerical variables. Then, clustering algorithms are designed for numerical and categorical datasets that are employed to produce corresponding clusters. On the other side, latent class clustering is a model-based clustering used for any type of data. The numbers of clusters base on the estimation of the probability model used. The best clustering method recommends LCC, which provides higher accuracy and the smallest standard deviation ratio. However, both LCC and cluster ensemble methods produce evaluation values that are not much different as the application method used potential village data in Bengkulu Province for clustering.

Keywords: clustering, cluster ensemble, LCC, mixed data, potential village data.

1. Pendahuluan

Penggerombolan merupakan analisis yang sering digunakan dalam mengelompokkan objek-objek berdasarkan ukuran kemiripan atau ketidakmiripan. Semakin mirip dua

* Received Jan 2020; Accepted Nov 2020; Published online on Nov 2020

objek maka semakin tinggi kemungkinan berada dalam suatu gerombol. Sebaliknya, semakin berbeda karakteristik suatu objek maka kemungkinan untuk berada dalam suatu gerombol semakin kecil. Penggerombolan dapat dikatakan ideal apabila tiap objek hanya masuk atau menjadi anggota dari salah satu gerombol sehingga tidak terjadi tumpang tindih (*overlapping*). Kemiripan dalam gerombol hanya berlaku dalam suatu gerombol dan antar gerombol berbeda. Ukuran kemiripan tersebut dapat dilihat berdasarkan jarak antar objek dan ukuran jarak dipengaruhi oleh jenis peubah yang digunakan. Apabila tidak dipertimbangkan pemilihan ukuran jarak yang digunakan terhadap peubah yang digunakan akan menyebabkan perhitungan jarak menjadi tidak valid.

Selain ukuran jarak yang diperhatikan dalam melakukan penggerombolan, ketepatan dalam menerapkan metode juga menjadi hal yang penting. Pada dasarnya semua metode menggunakan ukuran kesamaan atau ketidaksamaan antar objek (Johnson & Wichern, 2007). Akibatnya, penerapan metode juga memperhatikan jenis peubah yang digunakan karena objek yang akan dilakukan penggerombolan berasal dari peubah-peubah yang digunakan. Metode yang umumnya dikenal dalam proses pembentukan gerombol adalah metode hirarki (*hierarchical clustering*) dan metode tak berhirarki atau yang biasa dikenal dengan metode K-Means (*non-hierarchical clustering*). Kedua metode tersebut dapat digunakan apabila peubah bertipe numerik. Pada kenyataannya, data yang sering dijumpai ada yang mengandung peubah kategorik. Sehingga, memerlukan suatu metode penggerombolan yang tepat untuk diterapkan pada saat data yang digunakan berasal dari peubah yang berbeda yakni peubah campuran (numerik dan kategorik).

Beberapa metode yang telah dikembangkan dalam mengatasi permasalah penggerombolan pada peubah campuran adalah *Latent Class Cluster* (LCC), metode *Gower*, algoritma *K-Prototypes*, *Two Step Cluster Analysis* (TSCA), dan *cluster ensemble*. *Latent Class Cluster* (LCC) merupakan suatu teknik penggerombolan dengan peubah campuran yang diperkenalkan oleh Lazarsfeld pada tahun 1950 dengan menggunakan konsep pendugaan probabilitas posterior sebagai ukuran kemiripan objek. Sedangkan, pada tahun 1971 Gower memperkenalkan ukuran ketidakmiripan pada data campuran serta mengadopsi proses penggerombolan pada metode peubah numerik. Berbeda dengan metode *Gower*, algoritma *K-Prototypes* merupakan pengembangan dari metode K-Means pada penggunaan data dengan peubah berbeda. Selanjutnya, Chiu *et al*. (2001) mengembangkan metode TSCA dalam menangani penggerombolan dengan tipe skala pengukuran berbeda pula. Demikian juga dengan metode *cluster ensemble* yang diperkenalkan oleh Strehl & Ghosh (2002) dalam mengelompokkan peubah campuran dengan konsep menggabungkan dua metode berbeda sesuai jenis peubah dan pada tahun 2005 dikembangkan oleh He *et al*. (2005).

Pada penelitian ini akan dilakukan kajian terhadap dua metode diatas yaitu *cluster ensemble* dan LCC terhadap data dengan jenis peubah campuran. Selanjutnya, dari kedua metode tersebut untuk melihat ketepatan penggerombolan terbaik dapat dilihat berdasarkan rasio simpangan baku dalam geombol dan antar gerombol (Bunkers *et al*., 1996). Semakin kecil, nilai rasionya maka suatu metode dapat dikatakan melakukan penggerombolan dengan baik. Selain rasio keragaman, akan dilihat pula akurasi metode terhadap gerombol identitas yang sudah ditentukan dari awal. Hal ini

berguna untuk melihat apakah rasio terkecil yang diperoleh juga memberikan keakuratan terbesar.

Kedua metode diterapkan pada data simulasi hasil bangkitan sesuai sebaran masing-masing peubah untuk di perlihatkan kinerja metode terbaik. Metode terbaik yang terpilih diterapkan pada data riil, yaitu PODES (Potensi Desa) Provinsi Bengkulu untuk menggerombolokan desa/kelurahan yang ada di Provinsi Bengkulu.

2. Metodologi

2.1 Bahan dan Data

Terdapat dua jenis data yang digunakan pada penelitian ini yaitu data bangkitan (Tabel 1) dan data riil. Data bangkitan diperoleh menggunakan *software* R versi 3.5.2 dan menggunakan dua jenis peubah (Tabel 2 dan Tabel 3). Data tersebut digunakan untuk memperoleh metode penggerombolan terbaik yang kemudian akan digunakan pada penggerombolan data riil.

Tabel 1: Jenis data bangkitan pada simulasi

No data bangkitan	Ukuran gerombol	Nilai korelasi pada peubah numerik dalam gerombol	Ukuran data
1-3	1:1:1	0; 0.3; 0.9	150
4-6	1:1:2	0; 0.3; 0.9	150
7-9	1:2:3	0; 0.3; 0.9	120
10-12	1:1:1:1	0; 0.3; 0.9	100
13-15	1:1:2:2	0; 0.3; 0.9	120
16-18	1:2:3:4	0; 0.3; 0.9	100

Tabel 2: Jenis peubah pertama pada simulasi (V1) dengan k = 2

Numerik			Kategorik	
Peubah	Nilai tengah	Ragam	Peubah	Peluang
X1	5;20;30	1	X3	0.6;0.1;0.3
X2	10;25;35		X4	0.2;0.3;0.4;0.1

Tabel 3: Jenis peubah kedua pada simulasi (V2) dengan k = 4

Numerik			Kategorik	
Peubah	Nilai tengah	Ragam	Peubah	Peluang
X1	5;20;30;15	1	X3	0.25;0.25;0.25;0.25
X2	10;25;35;20		X4	0.20;0.30;0.20;0.30

Data riil yang digunakan pada penelitian ini adalah data sekunder hasil pendataan Potensi desa (PODES) Tahun 2018 yang dikumpulkan oleh Badan Pusat Statistik (BPS) dengan satuan pengamatan adalah seluruh desa dan kelurahan di Provinsi Bengkulu dengan total sebanyak 1514 desa/kelurahan dengan rincian sebanyak 1341 desa, 172 kelurahan dan 1 unit pemukiman transmigrasi di kabupaten Muko-Muko. Adapun pemilihan peubah penelitian didasarkan pada Buku Indeks Pembangunan Desa (IPD) 2018 yang meliputi 5 dimensi (Tabel 4).

Tabel 4: Daftar peubah penelitian berdasarkan pembagian dimensi IPD.

Dimensi	Peubah	Jenis
a. Pelayanan dasar mewakili aspek pelayanan guna mewujudkan bagian dari kebutuhan dasar	X1 : Jumlah keseluruhan sarana pendidikan	Numerik
	X2 : Jumlah keseluruhan sarana kesehatan	Numerik
	X3 : Jumlah keseluruhan tenaga kesehatan	Numerik
b. Kondisi infrastruktur mewakili kebutuhan dasar dan pendukung lainnya terkait dengan ketersediaan infrastruktur ekonomi, energi, air bersih dan sanitasi serta komunikas dan informasi.	X4 : Jumlah sarana dan prasarana ekonomi	Numerik
	X13 : Bahan bakar memasak sebagian besar keluarga	Kategorik 1. Gas Kota 2. LPG 3kg 3. LPG >3kg 4. Minyak tanah 5. Kayu bakar 6. Lainnya
	X5 : Jumlah keluarga pengguna listrik	Numerik
	X14 : Sumber air minum sebagian besar keluarga	Kategorik 1. Air Kemasan bermerk 2. Air isi ulang 3. Ledeng dengan meteran (PAM/PDAM) 4. Ledeng tanpa meteran 5. Sumur bor tanpa pompa 6. Sumur 7. Mata air 8. Sungai/danau/kolam/waduk/situ/embung/bendungan 9. Air hujan 10. Lainnya
	X15 : Sumber air mandi/cuci sebagian besar keluarga	Kategorik 1. Ledeng dengan meteran (PAM/PDAM) 2. Ledeng tanpa meteran 3. Sumur bor tanpa pompa 4. Sumur 5. Mata air 6. Sungai/danau/kolam/waduk/situ/embung/bendungan 7. Air hujan 8. Lainnya

Dimensi	Peubah	Jenis
	X16 : Kantor pos/pos pembantu/rumah pos	Kategorik 1. Beroprasi 2. Jarang beroperasi 3. Tidak beroperasi 4. Tidak ada
	X17 : Perusahaan/agen jasa ekspedisi (pengiriman barang/dokumen) swasta	Kategorik 1. Beroprasi 2. Jarang beroperasi 3. Tidak beroperasi 4. Tidak ada
	X18 : Fasilitas internet di kantor kepala desa/lurah	Kategorik 1. Berfungsi 2. Jarang berfungsi 3. Tidak berfungsi 4. Tidak ada
	X19 : Keberadaan warga yang menggunakan telepon seluler/handphone	Kategorik 1. Sebagian besar warga 2. Sebagian kecil warga 3. Tidak ada
c. Aksesbilitas/ Transportasi meliputi akses saranan dan prasarana transportasi yang mendukung kegiatan sosial ekonomi desa	X20 : Keberadaan angkutan umum	Kategorik 1. Ada, dengan trayek tetap 2. Ada, tanpa trayek tetap 3. Tidak ada angkutan umum
	X21 : Jenis permukaan jalan darat antar desa/kelurahan yang terluas	Kategorik 1. Aspal/beton 2. diperkeras (kerikil,batu,dll) 3. Tanah 4. lainnya
	X6 : Waktu tempuh dari kantor kepala desa/lurah ke kantor camat	Numerik
	X7 : Biaya transportasi dari kantor kepala desa/lurah ke kantor camat	Numerik
	X8 : Waktu tempuh dari kantor kepala desa/lurah ke kantor bupati/walikota	Numerik
	X9 : Biaya transportasi dari kantor kepala desa/lurah ke kantor bupati/walikota	Numerik

Dimensi	Peubah	Jenis
d. Pelayanan Umum merupakan upaya pemenuhan kebutuhan pelayanan kegiatan masyarakat. Peubah terpilih adalah terkait dengan fasilitas olahraga.	X22 : Fasilitas/lapangan olahraga Sepak bola	Kategorik 1. Ada, baik 2. Ada, rusak parah 3. Ada, rusak sedang 4. Tidak ada
	X23 : Fasilitas/lapangan olahraga bola voli	Kategorik 1. Ada, baik 2. Ada, rusak parah 3. Ada, rusak sedang 4. Tidak ada
	X24 : Fasilitas/lapangan olahraga bulu tangkis	Kategorik 1. Ada, baik 2. Ada, rusak parah 3. Ada, rusak sedang 4. Tidak ada
e. Penyelenggara Pemerintahan mewakili kinerja pemerintahan desa serta pendukungnya.	X10 : Umur Kepala desa/ Lurah	Numerik
	X11 : Umur Sekertaris desa/ Lurah	Numerik
	X12 : Jumlah aparatur pemerintahan desa	Numerik

2.2 Metode Penelitian

Adapun langkah analisis data yang dilakukan menggunakan *software* R versi 3.5.2. dalam penelitian ini adalah:

1. Menerapkan metode pada data bangkitan untuk dilakukan simulasi.
2. Penggerombolan *cluster ensemble* dilakukan dengan tahapan:
   a. Melakukan penggerombolan menggunakan algoritma *squeezer* pada jenis peubah kategorik.
   b. Melakukan penggerombolan pada jenis data numerik dengan metode Kmeans.
   c. Mengkombinasikan hasil gerombol kategorik terhadap hasil penggerombolan data numerik dan dianggap sebagai peubah kategorik sehingga dalam melakukan penggerombolan menggunakan algoritma *squeezer*.
   d. Menghitung rasio simpangan baku dalam gerombol dan antar gerombol dan akurasi.
3. Penggerombolan LCC dilakukan dengan tahapan:
   a. Menduga parameter probbilitas posterior dengan menggunakan metode EM
   b. Melakukan pengelompokan objek berdasarkan nilai probabilitas posterior yang tertinggi.
   c. Menghitung rasio simpangan baku dalam gerombol dan antar gerombol dan akurasi. Rasio simpangan baku untuk peubah numerik diberikan pada persamaan berikut:

$$S_W = \frac{1}{C}\sum_{c=1}^{C} S_c \quad \text{dan } S_B = \left[\frac{1}{C-1}\sum_{c=1}^{C}(\bar{x}_c - \bar{x})^2\right]^{\frac{1}{2}} \tag{1}$$

dengan,

S_W : nilai simpangan baku di dalam kelompok
S_B : nilai simpangan baku antar kelompok
S_c : nilai simpangan baku kelompok ke-*c*

$\bar{x}_c$: rata-rata kelompok ke-*c*
$\bar{x}$: rata-rata keseluruhan kelompok
C : jumlah kelompok yang terbentuk

Sedangkan S_w dan S_B untuk peubah dengan tipe kategorik adalah sebagai berikut:

$$S_W = \left[\frac{1}{n-C}\left[\sum_{c=1}^{C}\left(\frac{n_c}{2} - \frac{1}{2n_c}\sum_{k=1}^{K} n_{kc}^2\right)\right]\right]^{\frac{1}{2}} \text{ dan}$$

$$S_B = \left[\frac{1}{C-1}\left[\frac{1}{2}\left(\sum_{c=1}^{C}\frac{1}{n_c}\sum_{k=1}^{K} n_{kc}^2\right) - \frac{1}{2n}\sum_{k=1}^{K} n_k^2\right]\right]^{\frac{1}{2}} \quad (2)$$

dimana,

S_W : nilai simpangan baku di dalam kelompok data kategorik
S_B : nilai simpangan baku antar kelompok data kategorik
MSW : *Mean of squares within*
MSB : *Mean of squares between c*
SSW : *Sum of squares within*
SSB : *Sum of squares between*
C : jumlah kelompok yang terbentuk
n : banyaknya pengamatan
n_k : jumlah pengamatan dengan kategori ke-*k*
n_c : jumlah pengamatan pada kelompok ke-c,
n_{kc} : jumlah pengamatan dengan kategori ke-*k* dan kelompok ke-*c*

4. Membandingkan hasil penggerombolan antara metode cluster ensemble dan LCC dengan membandingkan nilai rasio simpangan baku dalam gerombol dan antar gerombol terkecil serta akurasi terbesar.
5. Menerapkan metode terbaik kedal data riil dan melakukan evaluasi terhadap kinerja metode terbaik terhadap keseluruhan metode sebelumnya tersebut dengan langkah sebagai berikut:
   a. Melakukan eksplorasi data pada setiap peubah.
   b. Melakukan penggerombolan terhadap metode terbaik yang dihasilkan pada kajian simulasi.
   c. Menghitung rasio simpangan baku dalam gerombol dan antar gerombol untuk menentukan gerombol optimal.

3. Hasil dan Pembahasan

3.1 Hasil Penerapan Metode pada Data Simulasi

Metode *cluster ensemble* dalam melakukan penggrombolan pada data simulasi dilakukan dengan memisahkan masing-masing peubah sesuai jenisnya dan menerapkan masing-masing peubah pada metode yang sesuai. Penggerombolan peubah numerik dilakukan dengan menggunakan metode Kmeans dan peubah kategorik menggunakan *algoritma squeezer*. Selanjutnya pemilihan algoritma *squeezer* didasarkan pada penelitian He *et al.* (2005). Hasil dari tahap penggerombolan masing-masing metode tersebut diasumsikan sebagai peubah kategorik untuk kembali dilakukan penggerombolan menggunakan algoritma *squeezer.*

Metode selanjutnya yang akan diterapkan dalam data simulasi adalah metode LCC, dimana proses penggerombolan dlakukan dengan mengasumsikan peubah numerik menyebar normal dan peubah kategorik menyebar multinomial. Pembentukan gerombol dengan melakukan pendugaan parameter probabilitas posterior sebagai batasan suatu objek masuk kedalam gerombol tertentu. Proses pendugaan dilakukan menggunakan algoritma *Expectation Maximization* (EM) diperoleh sampai pada peoses mengahasilkan dugaan yang konvergen.

Proses penerapan metode *cluster ensemble* dan LCC dilakukan dengan mendefinisikan banyaknya gerombol yang terbentuk pada metode sama dengan gerombol identitas yaitu 3 dan 4. Berdasarkan pengulangan sebanyak 1000 kali diperoleh hasil rasio simpangan baku dan nilai akurasi pada tiap-tiap jenis data bangkitan. Pemilihan rasio simpangan baku sebagai evaluasi kinerja metode dikarenakan semakin kecil nilai yang diperoleh memberikan makna bahwakualitas gerombol yang terbentuk semakin baik. Dengan kata lain, semakin kecil nilai rasio maka memberikan kemiripan yang semakin besar didalam gerombol dan menyebabkan semakin besar perbedaan tiap gerombol. Nilai rasio simpangan baku dihitung untuk masing-masing peubah sesuai dengan persamaan (1) dan (2). Selanjutnya dari rasio simpangan baku yang dihasilkan pada masing-masing peubah dilakukan perhitungan rata-ratanya. Proses simulasi dilakukan sebanyak 1000 kali untuk semua jenis bangkitan data, sehingga akan diperoleh 1000 nilai rasio simpangan baku. Kemudian, dari 1000 nilai tersebut dihitung kembali nilai rata-ratanya untuk digunakan sebagai perbandingan kinerja metode.

Selain rasio simpangan baku, nilai akurasi juga dipilih untuk mengevaluasi kinerja metode. Hal ini dikarenakan, data dibangkitkan sesuai karakteristik kemiripan masing-masing gerombol. Misalkan, untuk gerombol pertama pada peubah numerik pertama didefinisikan memiliki rata-rata sebesar 5 dan pada peubah numerik kedua sebesar 10 (Tabel 2). Berbeda dengan peubah numerik, peubah kategorik dibangkitkan sesuai kemunculan kategori demikian seterusnya sampai pada gerombol ketiga. Perhitungan akurasi dilakukan menggunakan tabel *Confusion Matrix* atau dikenal sebagai tabel ketepatan klasifikasi. Hasil penggerombolan kedua metode yang ditentukan banyaknya gerombol sesuai dengan gerombol identitas yang dibangkitkan akan dilakukan perhitungan ketepatan penggerombolannya menggunakan rumus akurasi pada *Confusion Matrix.* Nilai akurasi untuk satu jenis data simulasi akan diperoleh sebanyak 1000 sesuai banyaknya perulangan. Selanjutnya, seperti pada rasio simpangan baku untuk membandingkan dengan jenis data simulasi yang lain dilakukan perhitungan rata-rata terhadap nilai tersebut. Hasil proses simulasi untuk semua jenis data disajikan pada Tabel 5 berikut. Akurasi metode *cluster ensemble* dan Latent Class Clustering diberi simbol CE dan LCC. Rasio simpangan baku dihitung untuk gerombol identias sesuai data yang dibangkitkan dan kedua metode tersebut.

Berdasarkan Tabel 5 hasil akurasi dan rasio simpangan baku terhadap kenaikan korelasi tidak memberikan perubahan yang signifikan. Oleh karena itu, perlu adanya kajian lebih lanjut terkait hal ini. Namun, nilai yang dihasilkan kedua metode terhadap perbedaan banyaknya jenis gerombol dan rasio ukuran gerombol memberikan nilai yang konsisten. Secara keseluruhan, kedua metode memberikan nilai evaluasi yang tidak jauh berbeda. Hal ini memberikan makna bahwa kedua metode mampu melakukan penggerombolan pada peubah campuran dengan baik. Namun, apabila

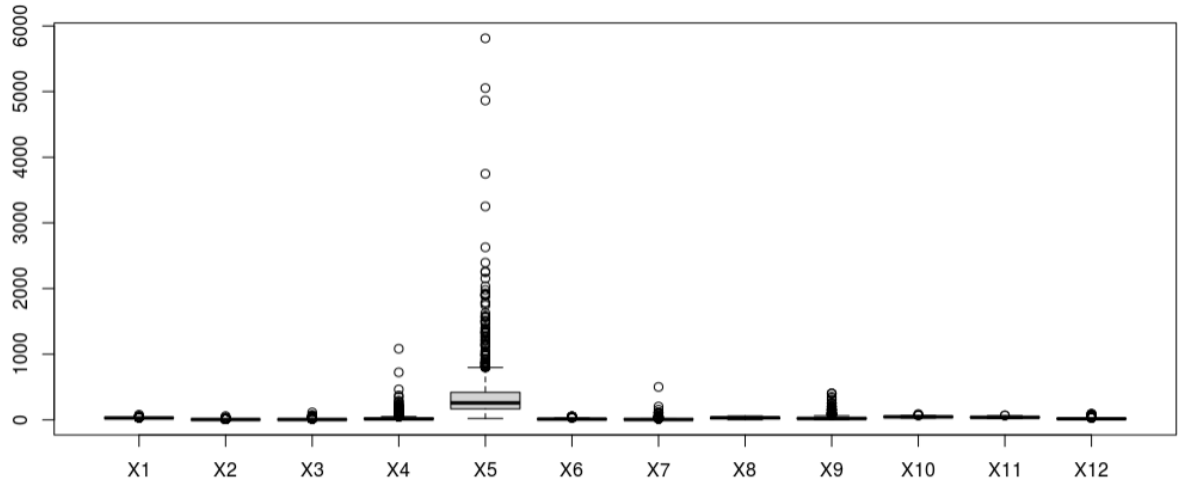
dilihat berdasarkan metode yang memberikan nilai yang terbaik paling banyak adalah metode LCC.

Tabel 5: Hasil Rataan Keseluruhan pada 1000 kali Ulangan

		Akurasi		Rasio Simpangan Baku		
Ukuran Gerombol	Nilai Korelasi	CE	LCC	ID	CE	LCC
1:1:1	0	98.07	98.23	0.539	0.549	0.532
	0.3	97.98	98.63	0.539	0.550	0.512
	0.9	96.27	98.54	0.539	0.558	0.523
1:1:2	0	95.87	98.42	0.632	0.645	0.637
	0.3	96.15	97.32	0.632	0.644	0.637
	0.9	95.65	98.21	0.632	0.647	0.638
1:2:3	0	93.21	96.31	0.609	0.643	0.621
	0.3	93.36	95.55	0.608	0.643	0.634
	0.9	95.65	95.60	0.632	0.647	0.649
1:1:1:1	0	96.77	97.82	0.674	0.686	0.678
	0.3	96.44	97.01	0.675	0.688	0.682
	0.9	95.83	97.39	0.674	0.690	0.676
1:1:2:2	0	99.04	99.51	0.663	0.670	0.667
	0.3	98.83	99.78	0.662	0.670	0.668
	0.9	98.63	99.01	0.663	0.670	0.667
1:2:3:4	0	97.22	98.86	0.722	0.746	0.734
	0.3	96.92	97.49	0.722	0.746	0.742
	0.9	96.36	97.85	0.711	0.727	0.723

3.2 Hasil Penerapan dan Evaluasi Metode Terbaik pada Data PODES

Data terapan yang digunakan adalah data PODES 2018 Provinsia Bengkulu, dengan 12 peubah numerik dan 12 peubah kategorik. Dua jenis peubah ini diasumsikan mengikuti sebaran normal dan multinomial. Hal ini dikarenakan data yang ada memiliki kaakteristik parameter yang dimiliki oleh kedua sebaran tersebut. Karakteritik pada peubah numerik disajikan dengan menggunakan diagram kotak garis (*boxplot*). Sedangkan untuk peubah kategorik disajikan dalam presentase masing-masing kategori yang muncul, serta ada beberapa kategori dalam peubah yang tidak muncul diakibatkan adanya desa yang tidak memuat kategori tersebut.



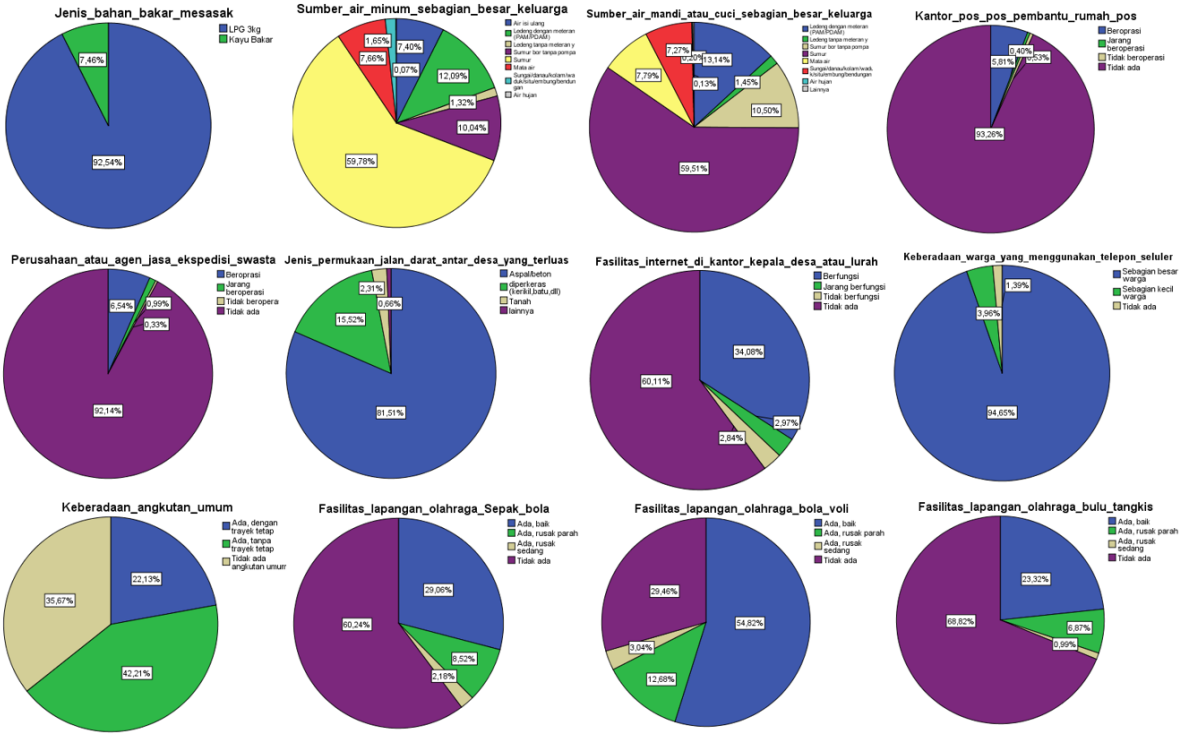
Gambar 1: Karakteristik peubah numerik data PODES.

Berdasarkan Gambar 1 peubah numerik selain X8 memiliki pencilan, dimana nilai pencilan yang diberikan terdapat dibagian atas. Pecilan tersebut merupakan nilai amatan yang lebih besar dari pada amatan lainnya. Pada peubah X5 yang merupakan banyaknya keluarga pengguna listrik memiliki sebaran nilai yang lebih besar diantara peubah yang lainnya.

Berdasarkan Gambar 2 untuk peubah kategorik diperoleh bahwa lebih dari 90% desa menggunakan LPG 3kg sebagai bahan bakar memasak dan sisanya menggunakan kayu bakar. Penggunaan air untuk minum, mandi atau mencuci lebih dari 50% desa menggunakan air sumur. Namun, melihat kondisi infrastruktur terkait keberadaan kantor pos dan agen jasa pengiriman swasta terlihat bahwa lebih dari 90% desa yang tidak memiliki infrastruktur tersebut. Hal ini berbanding terbalik dengan kondisi jalan yang sudah lebih dari 80% berbentuk aspal/beton. Ditemukan pula bahwa terdapat sekitar 60% tidak terdapat fasilitas internet di kantor desa/lurah. Lain halnya dengan kepemilikan telepon seluler bahwa terdapat lebih dari 90% desa sebagaian besar warganya telah memiliki telepon seluler. Selanjutnya, dilihat dari aspek transportasi hanya terdapat sekitar 42% desa yang memiliki angkutan umum tanpa trayek tetap. Terkait dengan keberadaan pelayanan umum dari aspek fasilitas olahraga sepak bola, bola voli dan bulu tangkis terdapat lebih dari 50% desa memiliki fasilitas tersebut.

Adapun hasil simulasi menunjukkan bahwa LCC adalah metode terbaik berdasarkan nilai rasio simpangan baku dan akurasi. Namun, pada LCC berlaku bahwa data pada peubah numerik yang digunakan harus memenuhi asumsi sebaran normal. Namu, pada saat dilakukan pengujian ternyata pada data PODES penelitian ini asumsi tersebut tidak terpenuhi. Berbagai transformasipun telah dilakukan dan memberikan hasil yang sama. Berdasarkan pertimbangan seblumnya bahwa *cluster ensemble* memberikan hasil yang tidak jauh berbada dengan LCC pada proses simulasi. Oleh karena itu, untuk melakukan pengerombolan pada data PODES diterapkan *cluster ensemble*. Tahap pertama yang dilakukan adalah menerapkan metode penggerombolan sesuai masing-masing jenis peubah, *Kmeans* untuk peubah numerik dan *Squeezer* untuk peubah kategorik. Tahap awal penggerombolan dilakukan dengan menentukan gerombol optimal. Pemilihan gerombol optimal peubah

numerik menggunakan *Package* R NbClust dengan mendefinisikan gerombol minimun adalah 2 dan gerombol maksimum adalah 15. Hasil pemilihan gerombol terbaik berdasarkan 26 indeks. Gerombol optimal yang terbentuk adalah sebanyak 2 gerombol. Pembentukan gerombol optimal pada peubah kategorik berdasarkan nilai *treshold*, pemilihan berdasarkan nilai rasio simpangan baku yang terkecil. Gerombol optimal yang terbentuk adalah 2 gerombol dengan nilai *treshold* 3.8-4.3 karena memberikan nlai Sw/Sb terkecil.



Gambar 2: Karakteristik peubah kategorik data PODES

Tahap kedua dalam *cluster ensemble* adalah mengasumsikan hasil penggerombolan optimum dari peubah numerik dan kategorik sebagai dua peubah kategorik baru untuk dilakukan penggerombolan menggunakan algoritma *Squeezer*. Hasil akhir pembentukan gerombol diperoleh berdasarkan nilai *treshold* yang hanya menghasilkan 3 jenis gerombol. Pemilihan gerombol optimal berdasarkan rasio simpangan baku terkecil dan gerombol optimal yang terbentuk yaitu 4 gerombol.

Visualisasi gerombol optimal untuk peubah numerik ditampilkan dalam bentuk *Chernoff face,* dengan memnghitung menggunakan rataan pada masing-masing peubah dalam gerombol. Deskripsi *Chernoff face* masing-masing peubah diberikan pada Tabel 6 berikut.

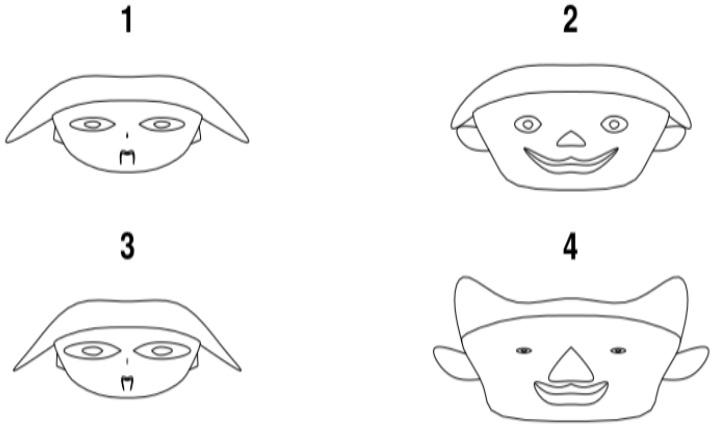
Tabel 6: Deskripsi Ciri Wajah pada Peubah Numerik

Peubah	Ciri Wajah	Keterangan
X1	*height of face/ width of nose*	Semakin tinggi bentuk wajah/semakin lebar bentuk hidung maka semakin banyak sarana pendidikan
X2	*width of face / width of ear*	Semakin lebar bentuk wajah/semakin lebar bentuk telinga maka semakin banyak sarana kesehatan
X3	*structure of face/ height of ear*	Semakin tidak teratur bentuk wajah/semakin tinggi bentuk telinga maka semakin banyak pengguna listrik
X4	*height of mouth*	Semakin tinggi bentuk mulut maka semakin banyak aparatur desa
X5	*width of mouth*	semakin lebar bentuk mulut semakin banyak tenaga kesehatan
X6	*smiling*	semakin melengkung keatas bentuk mulut semakin banyak sarana dan prasarana ekonomi
X7	*height of eyes*	semakin tinggi bentuk mata maka semakin lama waktu tempuh ke kantor camat
X8	*width of eyes*	semakin lebar bentuk mata maka semakin besar biaya ke kantor camat
X9	*height of hair*	semakin tinggi bentuk rambut maka semakin lama waktu tempuh ke kantor bupati
X10	*width of hair*	semakin lebar bentuk rambut maka semakin lama waktu tempuh ke kantor bupati
X11	*style of hair*	semakin melengkung keatas bentuk rambut maka semakin tinggi umur kades
X12	*height of nose*	semakin tinggi bentuk hidung maka semakin tinggi umur sekdes

Chernoff face dibuat dengan package *aplpack* pada program R dengan menggunakan fungsi *faces* (Gambar 3). Gerombol 1 dan 3 memiliki bentuk wajah yang mirip, sehingga dapat dikatakan bahwa dimensi IPD pada masing-masing peubah memiliki nilai rataan yang tidak jauh berbeda. Bentuk wajah pada gerombol 2 memiliki bentuk mulut yang paling lebar dan paling tersenyum diantara yang lainnya. Hal ini menunjukkan bahwa dimensi IPD pelayanan dasar pada aspek tenaga kesehatan dan dimensi kondisi infrastruktur aspek sarana dan prasarana ekonomi menunjukkan nilai rataan yang paling tinggi.

Bentuk ciri wajah pada Gerombol 2 menunjukkan bahwa untuk dimensi IPD pelayanan dasar aspek pendidikan, dimensi kondisi infrastruktur pada aspek penguna listrik dan dimensi penyelenggara pemerintahan untuk semua aspek memberikan nilai rataan yang paling tinggi. Hal tersebut ditunjukkan pada Gambar 3 yang menunjukkan ciri wajah yang jauh lebih menonjol dibandingkan gerombol yang lain sesuai deskripsi pada Tabel 6. Terdapat nilai rataan terkecil untuk beberapa peubah pada gerombol ini, yaitu pada dimensi aksesbilitas dan terlihat pada ciri wajah mata yang kecil serta pada tinggi dan lebar rambut yang lebih kecil dibandingkan gerombol 2. Namun hal ini, menunjukkan hal yang baik karena mengingat waktu tempuh dan biaya trasportasi

ke kantor camat dan bupati membutuhkan waktu dan biaya yang relatif lebih murah.



Gambar 3: Visualisasi *Chernoff face* pada Gerombol Terbaik Peubah Numerik

Pada peubah kategorik gerombol 1 lebih banyak didominasi oleh hampir semua kategori dalam peubah dibandingkan gerombol lain. Berdasarkan dimensi IPD untuk kondisi infrastruktur gerombol 1 memiliki karakteristik desa dengan jenis bahan bakar yang digunakan untuk memasak didominasi oleh penggunaan LPG 3kg meski masih terdapat sekitar 6% dari keseluruhan desa yang masih menggunakan kayu bakar. Sumber air minum dan mandi sebagian besar keluarga pada gerombol ini paling banyak berasal dari sumur. Selain itu, mayoritas tidak tersedia infrastruktur kantor pos serta agen pengiriman barang/dokumen lainnya. Sebagian besar warga menggunakan telepon seluler meskipun masih ada warga yang tidak menggunakan telepon seluler. Berdasarkan dimensi aksesbilitas/transportasi mayoritas terdapat angkutan umum tanpa trayek tetap dan jenis permukaan jalan mayoritas berbentuk aspal/beton. Berdasarkan dimensi pelayanan umum yang terdiri dari keberadaan fasilitas olahraga antara lain lapangan sepak bola, bola voli dan bulu tangkis masih banyak desa yang tidak memiliki fasilitas tersebut.

Hasil penggerombolan akhir menunjukkan bahwa terdapat kemiripan antara gerombol 1 dan 3 serta gerombol 2 dan 4. Hal ini dilihat berdasarkan intensitas kemunculan kategori pada masing-masing gerombol yang cenderung sama. Pada gerombol 3 memberikan tingkat IPD yang cenderung lebih tinggi dari pada gerombol 1. Dimensi IPD untuk gerombol 2 dan 4 kondisi infrastruktur memiliki karakteristik yang sama pada jenis bahan bakar memasak yaitu LPG. Dimensi pelayanan umum, transportasi/aksesbilitas dan kondisi infrastruktur terkait kepemilikan telepon seluler pada gerombol 4 hanya menampilkan satu jenis kategori yang menunjukkan IPD yang lebih baik dibandingkan gerombol laiinnya. Gerombol 2 sendiri menunjukkan IPD yang lebih baik daripada gerombol 1 dan 3.

Secara keseluruhan berdasarkan penggerombolan cluster *ensemble* diperoleh bahwa tingkat IPD yang terbaik ditunjukkan pada gerombol 4 yang memuat sebanyak 4 desa yaitu desa Cempaka Permai, Panorama, Pematang Gubernur dan Bentiring Permai.

4. Simpulan

Secara keseluruhan hasil metode terbaik yang diperoleh pada simulasi terhadap data bangkitan adalah LCC. Namun, *cluster ensemble* memberikan evaluasi nilai yang tidak jauh berbeda dengan LCC. Penerapan metode terbaik pada data PODES tidak dapat

dilakukan, karena peubah numerik pada data PODES tidak menunjukkan sebaran normal pada saat dilakukan pengujian. Oleh karena itu, menggunakan metode *cluster ensemble* sebagai alternatif lain yang dapat diterapkan kedalam data PODES. Gerombol optimal yang diperoleh *cluster ensemble* dipilih berdasarkan nilai Sw/Sb terkecil yaitu sebanyak 4 gerombol. Peubah terpilih dalam data PODES disusun berdasarkan dimensi IPD. Hasil visualisasi gerombol optimal pada *cluster ensemble* menunjukkan bahwa gerombol 4 memiliki IPD terbaik dengan memuat 4 desa yaitu Cempaka Permai, Panorama, Pematang Gubernur dan Bentiring Permai.

Daftar Pustaka

Bunkers, M. J., Miller Jr, J. R., & DeGaetano, A. T. (1996). Definition of climate regions in the Northern Plains using an objective cluster modification technique. *Journal of Climate*, *9*(1): 130–146.

Chiu, T., Fang, D., Chen, J., Wang, Y., & Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 263–268. San Francisco (US): ACM Press.

He, Z., Xu, X., & Deng, S. (2005). Clustering mixed numeric and categorical data: A cluster ensemble approach. *ArXiv Preprint Cs/0509011*, 1–14.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (Vol. 5). Prentice hall Upper Saddle River, NJ.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*(Dec): 583–617.