ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 9, Nomor 3, Tahun 2020, Halaman 263 - 272
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENERAPAN ANALISIS KLASTER *K*-MODES DENGAN VALIDASI DAVIES BOULDIN INDEX DALAM MENENTUKAN KARAKTERISTIK KANAL YOUTUBE DI INDONESIA
(Studi Kasus: 250 Kanal YouTube Indonesia Teratas Menurut *Socialblade*)

Ahmad Badruttamam[1], Sudarno[2], Di Asih I Maruddani[3]
[1, 2, 3]Departemen Statistika FSM Universitas Diponegoro
abdrtamam@gmail.com

ABSTRACT

YouTube is one of the most popular online platforms today. The popularity of YouTube has makes it an effective advertising medium. In April 2019, Socialblade released the top 250 YouTube channels in Indonesia based on their gradations with various characteristics. YouTube channel data will be grouped into several clusters to make it easier for advertisers to choose channels with characteristics as needed. The purpose of this study is to determine the best number of clusters and determine their characteristics. The method used is the k-Modes cluster analysis with values k = 3, 4, 5, ..., 8. The k-Modes method can group objects that have categorical type variables into relatively homogeneous groups. The best number of clusters (k) can be checked using the Davies Bouldin Index (DBI). Based on the analysis carried out, obtained the best number of six clusters with a Davies-Bouldin Index value of 1.080509. The most recommended cluster for advertising is cluster 6, which has grade A characteristics, gold title, and has an estimated annual income of 5 million USD < income ≤ 10 million USD.

Keywords: Youtube, Cluster Analysis, *k*-Modes, Categorical Data, Davies-Bouldin Index

1. PENDAHULUAN

YouTube merupakan salah satu *platform online* paling populer saat ini. Dikutip dari laman resminya, YouTube memiliki lebih dari satu miliar pengguna di seluruh dunia, angka ini mencakup hampir sepertiga dari jumlah pengguna internet secara keseluruhan. Popularitas yang dimiliki oleh YouTube membuatnya menjadi media iklan yang efektif. Salah satu hal yang menjadi masalah bagi pengiklan yaitu memilih kanal YouTube yang tepat untuk beriklan. Salah satu metode statistika yang tepat digunakan pada masalah tersebut adalah analisis klaster, karena analisis klaster dapat mengelompokkan objek ke dalam kelompok yang relatif homogen (Hair *et al*., 2014).

Metode analisis klaster yang sering digunakan adalah analisis klaster *k*-Means. Algoritma *k*-Means hanya bekerja terbatas pada *dataset* yang atributnya bertipe numerik. Algoritma klastering *k*-Modes merupakan pengembangan dari algoritma *k*-Means untuk data bertipe kategorik (Huang, 1997). Dalam penentuan klasternya dapat dibuat sebanyak dua, tiga, empat dan seterusnya, dimana setiap klaster mempunyai karakteristik yang sama. Jumlah klaster terbaik dapat diperiksa menggunakan *Davies-Bouldin Index* (DBI). Jumlah klaster yang dipilih adalah jumlah klaster yang memiliki nilai DBI terkecil (Hilmi, *et al.* 2015).

2. TINJAUAN PUSTAKA

2.1. YouTube

YouTube merupakan media sosial berbasis video yang didirikan pertama kali pada tanggal 14 Februari 2005 oleh Chad Hurley, Steve Chen, dan Jawed Karim. Ketiganya merupakan mantan karyawan PayPal. Kanal adalah tempat bagi pengguna agar dapat

mengelola konten video untuk audiens. Sebagai pemilik kanal, pengguna dapat menambahkan video, link, dan informasi tentang diri atau kanal agar dapat ditelusuri oleh pengunjung.

Socialblade membagi tipe kanal yang ada di YouTube menjadi beberapa kategori yaitu otomotif, komedi, pendidikan, hiburan, film, *gaming*, *howto* & *style*, musik, berita & politik, nonprofit & aktivisme, *people* & *blogs*, hewan peliharaan & binatang. Ditinjau dari segi kepemilikan, kanal YouTube terbagi atas personal dan non personal. Kanal personal merupakan kanal yang dimiliki dan dikelola oleh perorangan dan tidak berada dibawah naungan perusahaan atau instansi.

YouTube memberikan predikat dan plakat kepada kanal yang telah mencapai jumlah *subscriber* tertentu. *Silver Play Button* diberikan kepada kanal yang telah mencapai seratus ribu *subscriber*. *Gold Play Button* diberikan kepada kanal yang telah mencapai satu juta *subscriber*. *Diamond Play Button* diberikan kepada kanal yang telah mencapai sepuluh juta *subscriber*. *Ruby Play Button* diberikan kepada kanal yang telah mencapai lima puluh juta *subscriber*.

2.2. Socialblade

Dikutip dari laman resminya, Socialblade merupakan *platform* yang mengumpulkan data statistik dari berbagai macam media sosial seperti YouTube, Instagram, Twitter, dan lain sebagainya. Data tersebut kemudian diolah menjadi grafik dan bagan statistik yang memperlihatkan kemajuan dan pertumbuhan dari akun penggunanya.

Data statistik YouTube yang ditampilkan di *website* Socialblade adalah data statistik dari sebuah kanal, meliputi total unggahan, total *subscriber*, total tayangan (*view*), perkiraan pendapatan per tahun, serta data statistik lain yang menunjukkan pertumbuhan suatu kanal. Selain itu, Socialblade juga menampilkan *grade* dari suatu kanal. *Grade* yaitu standar pengukuran khusus yang dikembangkan oleh Socialblade yang memperhitungkan seberapa banyak *view* dan beberapa hal lain yang diperoleh suatu kanal untuk menentukan seberapa berpengaruh kanal tersebut. Beberapa tingkatan *grade* tersebut diurutkan dari yang tertinggi diantaranya adalah *grade* A++, A+, A, A-, B+, dan seterusnya

2.3. Analisis Klaster

Hair *et al.* (2014) mengemukakan bahwa analisis klaster adalah kumpulan dari beberapa teknik pengolahan data multivariat yang memiliki tujuan utama mengelompokkan objek-objek berdasarkan karakteristik yang dimilikinya. Hasil dari klaster yang terbentuk harus menunjukkan homogenitas internal yang tinggi dalam satu klaster dan heterogenitas yang tinggi antar klaster.

2.4. Algoritma *Clustering*

Menurut Hair *et al.* (2014) algoritma *clustering* terbagi ke dalam dua jenis, yaitu metode hirarki dan non hirarki. Metode hirarki dibagi menjadi dua yaitu, algoritma *agglomerative* (seperti *single-linkage, complete-linkage, average linkage,* metode *centroid* dan metode *ward*) dan algoritma *divisive*. Metode non hirarki misalnya *sequential threshold, parallel threshold* dan *optimization*.

Menurut Mattjik & Sumertajaya (2011) algoritma *clustering* hirarki digunakan untuk mengelompokkan objek secara terstruktur berdasarkan kemiripan sifatnya dan klaster yang diinginkan belum diketahui banyaknya. Menurut Johnson & Wichern (2002) algoritma *clustering* non hirarki digunakan untuk pengelompokan objek dimana banyaknya klaster yang akan dibentuk dapat ditentukan terlebih dahulu sebagai bagian dari prosedur pengelompokan.

2.5. *Clustering k*-Means

Algoritma dari metode *k*-Means sebagai berikut pertama tentukan besarnya *k* (yaitu banyaknya klaster, dan tentukan juga *centroid* di tiap klaster), kedua hitung jarak antara setiap objek dengan setiap *centroid*, ketiga hitung kembali rataan (*centroid*) untuk klaster yang baru terbentuk dan keempat ulangi langkah 2 sampai tidak ada lagi pemindahan objek antar klaster (Mattjik & Sumertajaya, 2011).

2.6. Variabel Kategorik

Agresti (1996) mengemukakan bahwa variabel kategorik merupakan salah satu skala pengukuran yang terdiri dari sejumlah kategori. Berdasarkan skala pengukuran, variabel kategorik dibagi menjadi:

1. Skala nominal, yaitu variabel kategorik yang tidak memiliki urutan nilai. Misalnya, variabel jenis musik yang disukai (*classical*, *country*, *folk*, *jazz*, *rock*), variabel jenis kelamin, dan lain sebagainya.
2. Skala ordinal, yaitu variabel kategorik yang memiliki urutan nilai. Misalnya, respon terhadap perawatan medis (sangat baik, baik, cukup, kurang) (Agresti, 1996).

2.7. *Clustering k*-Modes

Algoritma *clustering k*-Modes pertama kali diperkenalkan oleh Huang pada tahun 1997. Algoritma ini merupakan pengembangan dari algoritma *clustering k*-Means untuk mengelompokkan data kategorik. Algoritma *clustering k*-Means standar tidak dapat diaplikasikan untuk data kategorik. Hal ini dikarenakan fungsi jarak Euclidean dan penggunaan rata-rata untuk merepresentasikan pusat klaster. Langkah-langkah dalam algoritma *clustering k*-Modes sebagai berikut:

1. Memilih *k* modus awal sebagai titik pusat, satu untuk setiap klaster.
2. Menghitung jarak masing-masing data terhadap semua titik pusat klaster. Alokasikan setiap objek ke klaster terdekat menggunakan ukuran ketidaksamaan sederhana.
3. Setelah semua objek dialokasikan ke klaster, lakukan pengujian ulang perbedaan objek terhadap modus. Jika objek lebih mendekati klaster lain daripada klaster saat ini, maka alokasikan ulang objek ke klaster tersebut dan perbaharui modus kedua klaster.
4. Mengulangi langkah ketiga hingga tidak ada objek yang berubah klaster setelah dilakukan satu iterasi penuh terhadap seluruh data.

2.8. Ukuran Ketidaksamaan

Misalkan T_1, T_2 adalah dua objek yang dideskripsikan oleh *m* variabel kategorikal. Ukuran ketidaksamaan antara T_1 dan T_2 dapat didefinisikan sebagai total ketidakcocokan dari variabel kategorik yang sesuai dari kedua objek. Semakin kecil angka ketidakcocokan, semakin mirip kedua objek (Huang, 1997). Secara formal dirumuskan sebagai berikut.

$$d(T_1, T_2) = \sum_{j=1}^{m} \delta(x_{1j}, x_{2j}) \quad (1)$$

dimana

$$\delta(x_{1j}, x_{2j}) = \begin{cases} 0 & x_{1j} = x_{2j} \\ 1 & x_{1j} \neq x_{2j} \end{cases} \quad (2)$$

x_{1j} dan x_{2j} adalah nilai dari variabel ke-j pada objek T_1 dan T_2.

2.9. Modus Klaster

Pada *clustering k*-Modes, pusat klaster (*centroid)* diwakili oleh vektor modus dari variabel kategorik. Dalam statistika, modus dari sekumpulan nilai yaitu nilai yang paling sering muncul. Jika sekumpulan data memiliki *m* variabel kategorik, vektor modus *V* terdiri dari *m* nilai kategorik *(v_1, v_2, ..., v_m)*, yang masing-masing merupakan modus dari sebuah variabel. Vektor modus dari sebuah klaster meminimalkan jumlah jarak antar setiap obyek di dalam klaster dengan pusat klaster (Huang, 2009).

2.10. Indeks Davies-Bouldin

Menurut Permatadevi, *et al*. (2013) jika proses pengklasteran untuk masing-masing *k* selesai, maka untuk menentukan jumlah klaster yang terbaik dapat dilakukan penilaian menggunakan *Davies-Bouldin Index* (DBI). Pendekatan pengukuran ini bertujuan untuk memaksimalkan jarak antara klaster yang satu dengan yang lainnya dan pada waktu yang sama mencoba untuk meminimalkan jarak antara objek dalam sebuah klaster (Hilmi, *et al.*, 2015). Pengklasteran dengan jumlah klaster yang terbaik adalah pengklasteran yang memiliki nilai DBI minimum. Menurut Permatadevi, *et al.* (2013) nilai DBI dirumuskan pada persamaan (3).

$$DBI = \frac{1}{k}\sum_{a=1}^{k} R_a \tag{3}$$

dengan

$$R_a = \max_{b=1,\dots,k,a\neq b} R_{ab} \quad , \qquad R_{ab} = \frac{s_a + s_b}{d(V_a, V_b)} \tag{4}$$

dimana:

k = Jumlah klaster
R_{ab} = Ukuran kemiripan antara klaster ke-*a* dan klaster ke-*b*
s_a = Ukuran dispersi klaster ke-*a*, *a* = 1, 2, ..., *k*

$$s_a = \left[\frac{1}{n_a}\sum_{T_i \in C_a, i=1}^{n_a} d^2(T_i, V_a)\right]^{\frac{1}{2}}, \qquad d^2(T_i, V_a) = (d(T_i, V_a))^2 \tag{5}$$

dimana:

n_a = Banyaknya anggota klaster ke-*a*, *a* = 1, 2, ..., *k*
C_a = Klaster ke-*a*
T_i = Anggota ke-*i* pada klaster ke-*a*, *a* = 1, 2, ..., *k*

$d(T_i,V_a)$ adalah jarak dari anggota ke-*i* pada klaster ke-*a* (T_i) dengan *centroid* klaster ke-*a* (V_a) yang dihitung menggunakan ukuran ketidaksamaan pencocokan sederhana seperti pada persamaan (1) dan (2) sebagai berikut:

$$d(T_i, V_a) = \sum_{j=1}^{m} \delta\left(x_{ij}, v_{aj}\right) \tag{6}$$

dengan

$$\delta(x_{ij}, v_{aj}) = \begin{cases} 0 & x_{ij} = v_{aj} \\ 1 & x_{ij} \neq v_{aj} \end{cases} \tag{7}$$

dimana:

x_{ij} = nilai dari variabel ke-*j* pada T ke-*i*
v_{aj} = nilai ke-*j* pada *centroid* klaster ke-*a*
m = Jumlah variabel

$d(V_a,V_b)$ adalah jarak antara *centroid* klaster ke-*a* (V_a) dengan *centroid* klaster ke-*b* (V_b) yang dihitung menggunakan ukuran ketidaksamaan pencocokan sederhana sebagai berikut:

$$d(V_a,V_b) = \sum_{j=1}^{m} \delta(v_{aj},v_{bj}) \tag{8}$$

dengan

$$\delta(v_{aj},v_{bj}) = \begin{cases} 0 & v_{aj} = v_{bj} \\ 1 & v_{aj} \neq v_{bj} \end{cases} \tag{9}$$

dimana v_{bj} adalah nilai ke-*j* pada *centroid* klaster ke-*b*.

3. METODE PENELITIAN

3.1. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diambil dari *website* Socialblade.com untuk data 250 kanal YouTube teratas di Indonesia pada bulan April 2019. Pada data ini terdapat 3 data hilang, artinya informasi tentang data tidak lengkap dan tidak diikutsertakan dalam analisis sehingga dalam skripsi ini hanya digunakan 247 data dengan 5 variabel.

3.2. Variabel Penelitian

Tabel 1. Variabel Penelitian

Variabel	Nama Variabel
X_1	*Grade* Kanal
X_2	Predikat Kanal
X_3	Kategori Konten
X_4	Kategori Pemilik Kanal
X_5	Perkiraan Pendapatan per Tahun

3.3. Tahapan Analisis

Langkah-langkah yang dilakukan untuk menganalisis data pada penelitian ini adalah sebagai berikut:

1. Memasukkan data variabel.
2. Menentukan nilai *k* = 3, 4, 5, ..., 8.
3. Menentukan *centroid* (pusat klaster) secara acak dari masing-masing klaster
4. Menghitung jarak tiap objek terhadap *centroid* menggunakan ukuran ketidaksamaan sederhana.
5. Mengelompokkan objek berdasarkan jarak terdekat ke *centroid*.
6. Menentukan modus dari setiap variabel sebagai *centroid* klaster.
7. Menghitung ulang jarak tiap objek terhadap *centroid* baru menggunakan ukuran ketidaksamaan sederhana.
8. Mengelompokkan objek berdasarkan jarak terdekat ke *centroid*.
9. Apabila ada objek yang berpindah klaster, maka ulangi langkah ke enam sampai delapan hingga tidak ada objek yang berpindah klaster.
10. Memilih *k* terbaik menggunakan *Davies Bouldin Index.*
11. Membuat *profiling* dari masing-masing klaster.

4. HASIL DAN PEMBAHASAN

4.1. Metode *k*-Modes

Berikut ini disajikan ilustrasi langkah-langkah perhitungan manual proses pengklasteran dengan metode *k*-Modes. Proses pengklasteran dilakukan pada berbagai nilai *k* (*k* = 3, 4, 5, ..., 8).

a. Tentukan banyaknya klaster yang akan dibuat (*k*), misalnya nilai *k* = 3

b. Memilih 3 objek secara acak dan terpilih objek ke-50 , ke-157 , dan ke-242 sebagai *centroid* iterasi pertama yang disajikan pada tabel 2,

Tabel 2. Centroid Awal Klaster

Centroid **(V_a)**	**Objek ke-**	**X1**	**X2**	**X3**	**X4**	**X5**
V_1	50	2	2	14	2	3
V_2	157	1	3	6	1	3
V_3	242	1	3	12	2	1

c. Hitung jarak dari masing-masing objek dengan *centroid* iterasi pertama menggunakan ukuran ketidaksamaan pencocokan sederhana seperti pada persamaan (1) dan (2). Sebagai contoh jarak antara objek ke-1 terhadap *centroid* tiap klaster yaitu V_1, V_2, dan V_3 adalah sebagai berikut:

- Jarak objek ke-1 dengan *centroid* klaster 1
  $d(T_1, V_1) = \delta(x_{11}, v_{11}) + \delta(x_{12}, v_{12}) + \cdots + \delta(x_{15}, v_{15})$
  $d(T_1, V_1) = 1 + 1 + 1 + 0 + 1$
  $d(T_1, V_1) = 4$
  ⋮
  dan seterusnya sampai objek ke-247 ($d(T_{247},V_1)$)
- Jarak objek ke-1 dengan *centroid* klaster 2
  $d(T_1, V_2) = \delta(x_{11}, v_{21}) + \delta(x_{12}, v_{22}) + \cdots + \delta(x_{15}, v_{25})$
  $d(T_1, V_2) = 1 + 0 + 1 + 1 + 1$
  $d(T_1, V_2) = 4$
  ⋮
  dan seterusnya sampai objek ke-247 ($d(T_{247},V_2)$)
- Jarak objek ke-1 dengan *centroid* klaster 3
  $d(T_1, V_3) = \delta(x_{11}, v_{31}) + \delta(x_{12}, v_{32}) + \cdots + \delta(x_{15}, v_{35})$
  $d(T_1, V_3) = 1 + 0 + 1 + 0 + 1$
  $d(T_1, V_3) = 3$
  ⋮
  dan seterusnya sampai objek ke-247 ($d(T_{247},V_3)$)

d. Menentukan anggota dari masing-masing klaster berdasarkan jarak terdekat terhadap *centroid* klaster (V_a) dengan rumus jarak terdekat yaitu $min\{d(T_i,V_1), d(T_i,V_2), d(T_i,V_3)\}$.

e. Memperbarui *centroid* masing-masing klaster berdasarkan modus dari setiap variabel.

f. Hitung ulang jarak setiap objek terhadap *centroid* baru seperti pada langkah (c). Kemudian tentukan ulang anggota dari masing-masing klaster berdasarkan jarak terdekat terhadap *centroid* klaster terbaru seperti pada langkah (d).

Tabel 3. Hasil Pengklasteran k = 3 pada Iterasi Pertama dan Kedua

Objek ke-	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**Klaster**	
						Iterasi 1	**Iterasi 2**
1	3	3	1	2	5	3	3
2	3	3	1	2	5	3	3
3	3	3	1	2	4	3	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

31	2	3	14	1	3	1	2
32	2	2	1	2	1	1	1
33	2	3	1	2	4	1	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
245	1	2	6	1	2	2	2
246	1	3	1	2	2	3	3
247	1	2	14	2	3	1	1

Dari hasil pengklasteran pada iterasi kedua seperti pada Tabel 3 dapat diketahui bahwa ada objek yang berpindah klaster

g. Karena terjadi perpindahan objek, maka dilakukan pengulangan langkah (e) dan (f) hingga tidak ada lagi objek yang berpindah klaster.

Untuk pengklasteran k = 4, 5, ..., 8 dapat dilakukan dengan cara yang sama seperti langkah-langkah di atas.

4.2. Hasil Pengklasteran

Setelah proses pengklasteran mencapai iterasi maksimum, diperoleh hasil dari pengklasteran untuk masing-masing k yang disajikan pada Tabel 4.

Tabel 4. Hasil Pengklasteran k = 3, 4, 5, ..., 8

k	Klaster	Jumlah anggota (n_a)	*Centroid* (V_a)
3	1	109	1, 3, 1, 2, 3
	2	98	1, 3, 14, 1, 3
	3	40	1, 2, 8, 2, 2
4	1	98	1, 3, 1, 2, 3
	2	58	1, 3, 6, 1, 3
	3	36	1, 2, 8, 2, 2
	4	55	1, 2, 14, 1, 3
5	1	96	1, 3, 1, 2, 3
	2	53	1, 3, 6, 1, 3
	3	35	1, 2, 8, 2, 2
	4	48	1, 2, 14, 1, 3
	5	15	1, 1, 14, 1, 2
6	1	72	1, 3, 1, 2, 3
	2	94	1, 3, 14, 1, 3
	3	40	1, 2, 8, 2, 2
	4	25	1, 2, 1, 2, 3
	5	6	1, 1, 4, 1, 1
	6	10	3, 3, 1, 2, 4
7	1	74	1, 3, 1, 2, 3
	2	50	1, 3, 6, 1, 3
	3	31	1, 2, 8, 2, 2
	4	25	1, 2, 1, 2, 3
	5	37	1, 2, 14, 1, 2
	6	10	3, 3, 1, 2, 4
	7	20	1, 2, 14, 1, 3
8	1	35	2, 3, 1, 2, 3
	2	58	1, 3, 6, 1, 3
	3	42	1, 2, 1, 2, 2
	4	23	1, 2, 1, 2, 3
	5	38	1, 2, 14, 1, 2
	6	9	3, 3, 1, 2, 4
	7	20	1, 2, 14, 1, 3
	8	22	1, 3, 8, 2, 3

4.3. Penentuan Jumlah Klaster Terbaik

Dalam penentuan jumlah klaster terbaik digunakan metode Davies-Bouldin Index seperti yang dirumuskan pada persamaan (3). Semakin kecil nilai Davies-Bouldin Index akan memberikan hasil yang baik. Berikut ini disajikan contoh perhitungan nilai Davies-Bouldin Index untuk jumlah klaster k = 3.

Langkah pertama adalah menghitung jarak antar *centroid* klaster menggunakan persamaan (8) dan (9).

- Jarak antara *centroid* klaster 1 dan klaster 2

$d(V_1, V_2) = \delta(1,1) + \delta(3,3) + \delta(1,14) + \delta(2,1) + \delta(3,3)$
$d(V_1, V_2) = 0 + 0 + 1 + 1 + 0$
$d(V_1, V_2) = 2$

- Jarak antara *centroid* klaster 1 dan klaster 3
  $d(V_1, V_3) = \delta(1,1) + \delta(3,2) + \delta(1,8) + \delta(2,2) + \delta(3,2)$
  $d(V_1, V_3) = 0 + 1 + 1 + 0 + 1$
  $d(V_1, V_3) = 3$
- Jarak antara *centroid* klaster 2 dan klaster 3
  $d(V_2, V_3) = \delta(1,1) + \delta(3,2) + \delta(14,8) + \delta(1,2) + \delta(3,2)$
  $d(V_2, V_3) = 0 + 1 + 1 + 1 + 1$
  $d(V_2, V_3) = 4$

Kemudian akan dihitung jarak dari masing-masing anggota klaster terhadap *centroid* masing-masing klaster menggunakan persamaan (6) dan (7).

- Jarak antara masing-masing anggota klaster 1 terhadap *centroid* klaster 1
  $d(T_1, V_1) = \delta(3,1) + \delta(3,3) + \delta(1,1) + \delta(2,2) + \delta(5,3)$
  $d(T_1, V_1) = 1 + 0 + 0 + 0 + 1$
  $d(T_1, V_1) = 2$
  $\vdots$
  dan seterusnya sampai objek ke-109
- Jarak antara masing-masing anggota klaster 2 terhadap *centroid* klaster 2
  $d(T_1, V_2) = \delta(3,1) + \delta(3,3) + \delta(14,14) + \delta(1,1) + \delta(3,3)$
  $d(T_1, V_2) = 1 + 0 + 0 + 0 + 0$
  $d(T_1, V_2) = 1$
  $\vdots$
  dan seterusnya sampai objek ke-98
- Jarak antara masing-masing anggota klaster 3 terhadap *centroid* klaster 3
  $d(T_1, V_3) = \delta(2,1) + \delta(2,2) + \delta(8,8) + \delta(2,2) + \delta(2,2)$
  $d(T_1, V_3) = 1 + 0 + 0 + 0 + 0$
  $d(T_1, V_3) = 1$
  $\vdots$
  dan seterusnya sampai objek ke-40

Menghitung ukuran dispersi masing-masing klaster

- Ukuran dispersi klaster 1
  $$s_1 = \left[\frac{1}{109} \times (2^2 + 2^2 + 2^2 + \cdots + 2^2)\right]^{\frac{1}{2}}$$
  $s_1 = 1{,}889129$
- Ukuran dispersi klaster 2
  $$s_2 = \left[\frac{1}{98} \times (1^2 + 2^2 + 2^2 + \cdots + 2^2)\right]^{\frac{1}{2}}$$
  $s_2 = 1{,}622545$
- Ukuran dispersi klaster 3
  $$s_3 = \left[\frac{1}{40} \times (1^2 + 1^2 + 2^2 + \cdots + 2^2)\right]^{\frac{1}{2}}$$
  $s_3 = 1{,}440486$

Menghitung ukuran kemiripan antar klaster menggunakan persamaan (4) sebagai berikut

- Ukuran kemiripan antara klaster 1 dan klaster 2

$$R_{1,2} = \frac{s_1 + s_2}{d(V_1, V_2)}$$

$$R_{1,2} = \frac{1{,}889129 + 1{,}622545}{2} = 1{,}755837$$

- Ukuran kemiripan antara klaster 1 dan klaster 3

$$R_{1,3} = \frac{s_1 + s_3}{d(V_1, V_3)}$$

$$R_{1,3} = \frac{1{,}889129 + 1{,}440486}{3} = 1{,}109872$$

- Ukuran kemiripan antara klaster 2 dan klaster 3

$$R_{2,3} = \frac{s_2 + s_3}{d(V_2, V_3)}$$

$$R_{2,3} = \frac{1{,}622545 + 1{,}440486}{4} = 0{,}765758$$

Jika dibuat dalam bentuk matriks, maka

$$\boldsymbol{R} = \begin{bmatrix} - & 1{,}755837 & 1{,}109872 \\ 1{,}755837 & - & 0{,}765758 \\ 1{,}109872 & 0{,}765758 & - \end{bmatrix}$$

sehingga nilai Davies-Bouldin Index untuk *k* = 3 adalah:

$$\text{DBI} = \frac{1}{3}\sum_{a=1}^{3} \text{R}_a = \frac{1}{3}(1{,}755837 + 1{,}755837 + 1{,}109872) = 1{,}540515$$

Hasil perhitungan nilai Davies-Bouldin Index untuk *k* = 3, 4, 5, ..., 8 secara ringkas disajikan pada Tabel 5.

Tabel 5. Nilai Davies-Bouldin Index untuk k = 3, 4, 5, ..., 8

k	**DBI**
3	1,540515
4	1,414286
5	1,290032
6	1,080509
7	1,192608
8	1,144995

Berdasarkan Tabel 5 diperoleh nilai Davies-Bouldin Index terkecil yaitu 1,080509 pada *k* = 6, sehingga pada percobaan ini jumlah klaster terbaik adalah enam klaster.

4.4. Interpretasi dan Profiling Hasil Klaster untuk *k*=6

Tabel 6. Karakteristik Masing-masing Klaster untuk *k*=6

Klaster	**Variabel**				
	X_1	**X_2**	**X_3**	**X_4**	**X_5**
1	B+	*Gold*	*Entertainment*	Non personal	1 juta USD < pendapatan ≤ 5 juta USD
2	B+	*Gold*	Lainnya	Personal	1 juta USD < pendapatan ≤ 5 juta USD
3	B+	*Silver*	Musik	Non personal	0 < pendapatan ≤ 1 juta USD
4	B+	*Silver*	*Entertainment*	Non personal	1 juta USD < pendapatan ≤ 5 juta USD
5	B+	Tanpa Predikat	Edukasi	Personal	Tanpa pendapatan
6	A	*Gold*	*Entertainment*	Non personal	5 juta USD < pendapatan ≤ 10 juta USD

Apabila keenam klaster dibandingkan, maka klaster yang paling direkomendasikan untuk beriklan adalah klaster 6. Karena ditinjau dari segi *grade*, *grade* A merupakan *grade*

tertinggi. Semakin tinggi tingkat *grade* suatu kanal, semakin besar jumlah *view* yang diperoleh. Ditinjau dari segi predikat kanal, predikat *gold* menjadi predikat tertinggi yang menunjukkan jumlah *subscriber* kanal tersebut telah mencapai satu juta *subscriber*. Ditinjau dari segi perkiraan pendapatan per tahun, kategori pendapatan 5 juta USD < pendapatan ≤ 10 juta USD merupakan kategori pendapatan terbesar diantara keenam klaster.

4. KESIMPULAN

Berdasarkan uraian hasil analisis dan pembahasan, maka diperoleh kesimpulan sebagai berikut:

1. Jumlah klaster terbaik yang dihasilkan menggunakan metode *k*-Modes adalah enam klaster ($k = 6$) yang mempunyai nilai validitas Davies-Bouldin Index paling kecil yaitu sebesar 1,080509. Klaster yang terbentuk memiliki homogenitas internal yang tinggi dalam satu klaster dan heterogenitas yang tinggi antar klaster.
2. Terbentuk enam klaster dengan banyaknya anggota dari klaster 1 sebanyak 72 kanal, klaster 2 sebanyak 94 kanal, klaster 3 sebanyak 40 kanal, klaster 4 sebanyak 25 kanal, klaster 5 sebanyak 6 kanal, klaster 6 sebanyak 10 kanal.
3. Klaster yang paling direkomendasikan untuk beriklan adalah klaster 6 dengan karakteristik *grade* A, berpredikat *gold*, serta memiliki perkiraan pendapatan per tahun sebesar 5 juta USD < pendapatan ≤ 10 juta USD.

DAFTAR PUSTAKA

Agresti, A., 1996. *An Introduction to Categorical Data Analiysis.* Canada: John Wiley & Sons, Inc..

Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E., 2014. *Multivariate Data Analysis 7th*. USA: Pearson.

Hilmi, M., N., 2015. Pemetaan Preferensi Mahasiswa Baru dalam Memilih Jurusan Menggunakan Artificial Neural Network (ANN) dengan Algoritma Self Organizing Maps (SOM). *Jurnal Gaussian* Vol. 4, No. 1: Hal. 53-60

Huang, J. Z., 2009. *Clustering Categorical Data with k-Modes,* Hong Kong: IGI Global.

Huang, Z., 1997. *A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining.* Canberra: Cooperative Research Centre for Advanced Computational Systems.

Johnson, R. A. & Wichern, D. W., 2002. *Applied Multivariate Statistical Analysis 5th*. New Jersey: Pearson.

Mattjik, A. A. & Sumertajaya, I. M., 2011. *Sidik Peubah Ganda dengan Menggunakan SAS.* Bogor: IPB Press.

Permatadevi, M. A., Hendrawan, R. A., & Hafidz, I., 2013. Karakteristik Pelanggan Telepon Kabel Menggunakan Clustering SOM dan K-Means untuk Mengurangi Kesalahan Klasifikasi Pelanggan Perusahaan Telekomunikasi. *Jurnal Teknik Pomits* Vol. 1, No. 1 Hal. 1-6.