JSA
Jurnal Statistika dan Aplikasinya

Volume 3 Issue 1,
June 2019

e-ISSN: 2620-8369

Received: 11 Februari 2019
Revised: 5 Maret 2019
Accepted: 18 Maret 2019
Published: 28 Juni 2019

PENGELOMPOKAN PENGGUNA INTERNET DENGAN METODE *K-MEANS CLUSTERING*

Disi Amalia P.[1, a)], Bagus Sumargo[2, b)]

[1]*Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta*
[2] *Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta*

Email: [a)]disipramudita1997@gmail.com, [b)]bagussumargo63@gmail.com

Abstract

Digital disparities still exists between internet users in urban and rural areas of Indonesia, therefore it is necessary to classify provinces based on the percentage of internet users. The grouping method used is the K-Means Clustering method. The purpose of this study is to group provinces based on internet users into three clusters to find out which provinces have good infrastructure so that they can support internet use. The average percentage of internet users in each cluster 1, cluster 2, and cluster 3 is 40.91%; 72.27%; and 57.60%. There needs to be an equal distribution of infrastructure that supports internet use in Indonesian society.

Keywords: K-Means Clustering, Clustering, Internet users.

Abstrak

Kesenjangan digital masih terjadi antara pengguna internet di wilayah perkotaan dan pedesaan Indonesia, oleh karena itu, perlu dilakukan upaya untuk mengelompokan provinsi berdasarkan persentase pengguna internet. Metode pengelompokan yang digunakan adalah metode *K-Means Cluster*ing. Tujuan dari studi ini adalah mengelompokan provinsi berdasarkan pengguna internet menjadi tiga *cluster* untuk mengetahui provinsi-provinsi mana yang memiliki infrastruktur yang baik, sehingga dapat mendukung penggunaan internet. Rata-rata persentase jumlah pengguna internet masing-masing *cluster* 1, *cluster* 2, dan *cluster* 3 adalah 40,91%; 72,27%; dan 57,60%. Perlu adanya pemerataan infrastruktur yang mendukung penggunaan internet pada masyarakat Indonesia.

Kata-kata kunci *K-Means Cluster*ing, 3 *Cluster*, Pengguna Internet.

PENDAHULUAN

Pada era globalisasi saat ini, ilmu pengetahuan dan teknologi berkembang dengan begitu pesat. Salah satu akibat dari berkembangnya teknologi yang semakin canggih adalah meluasnya penggunaan internet. Pengguna internet di Indonesia semakin meningkat setiap tahunnya. Walaupun pengguna

internet di Indonesia semakin meningkat, namun masih ada kesenjangan digital yang kuat antara anak dan remaja yang tinggal di wilayah perkotaan (lebih sejahtera) di Indonesia, dengan mereka yang tinggal di wilayah pedesaan (dan kurang sejahtera). Di daerah perkotaan, hanya 13 persen dari anak dan remaja yang tidak menggunakan internet, sementara di daerah pedesaan ada 87 persen anak dan remaja yang tidak memakai internet (Panji, 2014).

Untuk itu perlu mengelompokkan provinsi berdasarkan persentase siswa umur 5-24 tahun yang mengakses internet selama 3 bulan terakhir menurut jenjang pendidikan. Pengelompokan dilakukan untuk mengetahui provinsi-provinsi mana yang memiliki infrastruktur baik untuk mendukung penggunaan internet dan provinsi-provinsi mana yang perlu dilakukan perbaikan infrastruktur untuk mendukung penggunaan internet pada masyarakat. Salah satu metode yang dapat digunakan untuk melakukan proses pengelompokan adalah analisis *cluster*.

Pada dasarnya, analisis *cluster* adalah metode analisis data yang termasuk dalam analisis statistik multivariat metode interdependensi, karena itu tujuan analisis *cluster* tidak untuk menghubungkan ataupun membedakan antar beberapa variabel. Analisis *cluster* terbagi menjadi dua metode yaitu, metode hierarki dan metode non-hierarki. Pada penelitian ini akan digunakan metode *k-means clustering* yang termasuk dalam metode non-hierarki untuk menganalisis data persentase siswa umur 5-24 tahun yang mengakses internet selama 3 bulan terakhir menurut jenjang pendidikan di Indonesia. *K-Means cluster*ing mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien (Arai, 2007).

Mengacu pada hal di atas maka akan dilakukan pengelompokan provinsi berdasarkan Persentase Pengguna Internet menurut Jenjang Pendidikan dengan Metode *K-Means Cluster*ing dan bantuan aplikasi *software* IBM SPSS 25.

METODE

Analisis *Cluster*

Analisis multivariat adalah suatu metode statistika yang digunakan untuk mengolah atau menganalisis data yang memiliki banyak variabel. Tujuan analisis multivariat adalah untuk mencari pengaruh variabel-variabel tersebut terhadap suatu objek. Analisis multivariat terbagi menjadi 2 kategori yaitu metode dependensi dan metode interdependensi.

Analisis *cluster* merupakan salah satu analisis multivariat yang mengelompokan objek-objek atau data ke dalam beberapa kelompok dimana setiap objek dalam satu kelompok memiliki karakteristik yang sama. Suatu *cluster* dikatakan *cluster* yang baik jika memiliki ciri-ciri sebagai berikut:

1. Homogenitas internal (*within-cluster*), yaitu memiliki tingkat kesamaan yang tinggi antar anggota dalam satu *cluster*.
2. Heterogenitas eksternal (*between-cluster*), yaitu memiliki tingkat perbedaan yang tinggi antar *cluster* yang satu dengan *cluster* yang lain.

Adapun langkah-langkah untuk mengelompokan data dengan menggunakan analisis *cluster* adalah sebagai berikut.

Melakukan Proses Standardisasi Data

Dalam analisis *cluster* terkadang data yang didapat memiliki satuan yang beranekaragam. Jika terdapat satuan yang berbeda dalam data, maka perlu dilakukan proses standardisasi data dengan mengubahnya ke bentuk z-*score*. Standardisasi data dapat dilakukan dengan rumus berikut:

$$z = \frac{x - \overline{X}}{\sigma} \tag{1}$$

Keterangan
z : Nilai data setelah di standardisasi
x : Nilai data asli

$\overline{X}$: Nilai rata - rata keseluruhan keseluruhan
σ : Simpangan baku

Menentukan Ukuran Kemiripan antar Data/Objek

Dalam analisis *cluster* diperlukan beberapa ukuran untuk mengetahui seberapa mirip objek-objek yang akan dikelompokan ke dalam *cluster* yang sama. Terdapat tiga metode yang digunakan dalam mengukur kemiripan / kesamaan antar objek yaitu ukuran korelasi, asosiasi dan jarak. Dalam ukuran jarak, jarak yang kecil menunjukan tingginya tingkat kesamaan antar objek. Dalam mengukur jarak antara dua objek, dapat digunakan ukuran jarak *Euclidean*. Berikut merupakan rumus mencari jarak *Euclidean*:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \quad (2)$$

Keterangan
d_{ij} : Jarak antara objek ke-i dengan objek ke-j
i : 1,2,3, ... ,n
j : 1,2,3, ... ,n
k : 1,2,3, ... ,p
n : Banyaknya objek
p : Banyaknya peubah atau variabel
x_{ik} : Nilai/data objek ke-i variabel ke-k
x_{jk} : Nilai/data objek ke-j variabel ke-k

Melakukan Proses *Clustering* dengan Metode yang Dipilih

Terdapat dua metode pengelompokan yang dapat digunakan dalam analisis *cluster*, yaitu sebagai berikut.

1. Metode Hirarki

   Pengelompokan dengan metode hirarki dimulai dengan mengelompokan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian proses pengelompokan diteruskan ke objek lainnya yang mempunyai kedekatan kedua. Demikian seterusnya sampai *cluster* membentuk tingkatan (hirarki). Hasil pengelompokan metode hirarki membentuk sebuah *dendogram* yang digunakan untuk membantu menjelaskan proses hirarki tersebut. Metode hirarki terbagi menjadi dua, yaitu sebagai berikut.

   a. Metode *Agglomerative* (Penggabungan)

      Pengelompokan metode *Agglomerative* dimulai dengan menganggap setiap objek sebagai sebuah *cluster*. Kemudian dua objek dengan jarak terdekat digabungakan menjadi satu *cluster* baru. Selanjutnya objek ketiga akan bergabung dengan *cluster* yang ada atau bergabung dengan objek lain membentuk *cluster* baru dan demikian seterusnya, dimana kedekatan jarak antar objek tetap diperhitungkan. Proses akan berlanjut hingga akhirnya terbentuk satu *cluster* yang terdiri dari keseluruhan objek. Metode *Agglomerative* terbagi menjadi beberapa jenis yaitu metode *single linkage*, metode *complete linkage*, metode *average linkage*, metode *ward*, dan metode *centroid*.

   b. Metode *Divisive* (Pemecahan)

      Pengelompokan metode *Divisive* dimulai dari satu *cluster* besar yang terdiri dari semua objek. Kemudian dua objek dengan jarak terjauh atau nilai ketidakmiripannya tertinggi akan dipisahkan dan membentuk *cluster* yang lebih kecil dan demikian seterusnya. Proses ini akan berlanjut hingga setiap objek menjadi *cluster* tersendiri.

2. Metode Non-Hirarki

Berbeda dengan metode hirarki, proses metode non-hirarki dimulai dengan menentukan jumlah *cluster* yang diinginkan terlebih dahulu. Setelah jumlah *cluster* ditentukan, maka proses pengelompokan dilakukan tanpa mengikuti proses hirarki. Metode ini biasanya disebut dengan *K-means clustering*.

Melakukan Interpretasi terhadap Hasil *Clustering*

Tahap ini dimulai dengan menganalisis variabel-variabel yang membedakan antar setiap *cluster*. Analisis dilakukan dengan melihat nilai signifikansi (sig.) dan nilai *F* pada tabel ANOVA. Apabila nilai sig. > 0,05, maka tidak ada perbedaan yang berarti antara masing-masing *cluster* dan jika nilai sig. < 0,05 maka ada perbedaan yang berarti antara masing-masing *cluster* yang berhubungan dengan variabel tersebut. Kemudian semakin besar nilai *F* maka semakin besar pula perbedaan antara masing-masing *cluster* yang berhubungan dengan variabel tersebut.

Interpretasi hasil *clustering* dapat dilakukan berdasarkan *output final cluster centers*. Jika nilai *output final cluster centers* berbentuk *z-score* maka terdapat ketentuan sebagai berikut.

1. Nilai negatif (-) artinya rata-rata variabel dalam *cluster* berada di bawah rata-rata populasi
2. Nilai positif (+) artinya rata-rata variabel dalam *cluster* berada di atas rata-rata populasi

Biasanya, interpretasi hasil *clustering* juga dapat dilakukan berdasarkan nilai rata-rata keseluruhan variabel pada tiap *cluster* yang terbentuk.

K-Means Clustering

Metode *K-Means* pertama kali diperkenalkan oleh James MacQueen pada tahun 1976. *K-Means Clustering* adalah metode pengelompokan non-hirarki yang mengelompokan data/objek ke dalam *cluster-cluster*. *K-Means Clustering* merupakan metode pengelompokan yang banyak digunakan karena sederhana dan mudah diimplementasikan. *K-Means Cluster*ing digunakan sebagai alternatif metode *cluster* untuk data berukuran besar karena memiliki kecepatan yang lebih tinggi dibandingkan metode hirarki (Sitepu, 2011).

Tujuan dari *K-Means clustering* adalah mengelompokkan data/objek dengan memaksimalkan kemiripan data dalam satu *cluster* dan meminimalkan kemiripan data antar *cluster*. Ukuran kemiripan yang digunakan dalam *k-means clustering* adalah ukuran jarak. Adapun langkah-langkah dari algoritma *k-means* adalah sebagai berikut.

1. Menentukan banyak *k-cluster* yang ingin dibentuk.
2. Membangkitkan nilai random untuk pusat *cluster* (*centroid*) awal sebanyak *k-cluster*.
3. Menghitung jarak setiap data terhadap masing-masing *centroid*. Untuk menghitung jarak data ke *centroid* dapat menggunakan ukuran jarak seperti jarak *euclidean*.
4. Mengelompokan setiap data berdasarkan jarak terkecil/terdekat antara data dengan *centroid*.
5. Mengupdate nilai *centroid*. Nilai *centroid* baru diperoleh dari rata-rata dari data pada setiap *cluster* dengan menggunakan rumus di bawah.

$$C_{ij} = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{kj} \tag{3}$$

Keterangan

C_{ij} : pusat *cluster* (*centroid*) dari *cluster* ke-i pada variabel ke-j

n_i : jumlah data yang menjadi anggota *cluster* ke-i

x_{kj} : nilai data ke-k yang ada di dalam *cluster* pada variabel ke-j

i, k : indeks dari *cluster*

j : indeks dari variabel

6. Melakukan pengulangan dari langkah 3 sampai 5 hingga anggota tiap *cluster* atau *centroid* tidak ada yang berubah.

HASIL DAN PEMBAHASAN

Data

Data yang digunakan dalam penelitian ini diperoleh melalui bagian pusat data dan sarana informatika (PDSI) Kementerian Komunikasi dan Informatika Republik Indonesia yang berupa data persentase siswa umur 5-24 tahun yang mengakses internet selama 3 Bulan terakhir menurut Provinsi dan Jenjang Pendidikan, 2017 (Perkotaan + Perdesaan). Data-data tersebut berisi data persentase jumlah pengguna internet berpendidikan SD, SMP, SMA dan Perguruan tinggi yang akan dijadikan variabel dalam penelitian ini. Data tersebut disajikan dalam tabel berikut.

Proses Analisis Cluster

Beberapa langkah untuk mengelompokan data ke dalam beberapa *cluster*, yaitu sebagai berikut.

TABEL 1. Data Jumlah Persentase Pengguna Internet Berpendidikan SD, SMP, SMA, dan Perguruan Tinggi (PT) per Provinsi 2017

No	Provinsi	SD	SMP	SMA	PT
1	Aceh	3,33	24,37	52,61	79,03
2	Sumatera Utara	8,79	44,92	74,60	90,18
3	Sumatera Barat	11,27	53,90	81,64	88,56
4	Riau	10,21	50,70	75,29	88,74
5	Jambi	9,00	47,27	78,36	89,00
6	Sumatera Selatan	8,29	43,49	75,91	93,30
7	Bengkulu	8,96	45,76	73,74	84,02
8	Lampung	6,77	42,24	79,80	83,48
9	Kep. Bangka Belitung	11,71	47,89	83,13	91,45
10	Kepulauan Riau	18,75	68,86	88,37	97,92
11	DKI Jakarta	28,74	76,56	93,03	99,77
12	Jawa Barat	13,35	64,42	88,84	93,64
13	Jawa Tengah	13,94	66,93	90,16	93,24
14	DI Yogyakarta	25,73	82,66	95,97	98,68
15	Jawa Timur	15,09	66,78	86,45	94,99
16	Banten	12,52	55,22	85,40	94,27
17	Bali	22,34	72,04	91,72	95,14
18	Nusa Tenggara Barat	4,29	32,60	66,71	78,39
19	Nusa Tenggara Timur	2,92	19,74	51,60	83,48
20	Kalimantan Barat	8,06	39,52	67,74	86,46
21	Kalimantan Tengah	11,73	48,14	76,19	76,44
22	Kalimantan Selatan	16,23	61,92	84,43	90,65
23	Kalimantan Timur	17,26	59,48	85,68	93,90
24	Kalimantan Utara	11,38	46,06	85,49	93,45
25	Sulawesi Utara	13,88	50,35	77,20	90,28
26	Sulawesi Tengah	6,03	39,80	72,01	84,10
27	Sulawesi Selatan	10,83	51,13	81,01	92,73
28	Sulawesi Tenggara	5,48	34,76	71,41	88,43

29	Gorontalo	7,63	45,90	77,49	90,31
30	Sulawesi Barat	3,60	26,53	63,35	81,23
31	Maluku	6,54	37,57	56,86	78,80
32	Maluku Utara	3,17	21,85	39,83	66,09
33	Papua Barat	6,45	33,61	62,05	75,91
34	Papua	3,87	20,22	44,03	78,45

Proses Standardisasi Data

Hal yang perlu diperhatikan dalam analisis *cluster* adalah satuan dari data-data penelitian, apakah memiliki keanekaragaman atau tidak. Data-data pada Tabel 1 belum memiliki satuan yang seragam. Oleh karena itu, perlu dilakukan proses standardisasi data terlebih dahulu ke bentuk z-*score*. Untuk mengetahui nilai rata-rata dan simpangan baku, dapat dilihat dari tampilan *output* SPSS berikut.

TABEL 2. *Descriptive Statistics*

	N	Minimum	Maximum	Mean	Std. Deviation
SD	34	2,92	28,74	10,8276	6,32304
SMP	34	19,74	82,66	47,7409	16,20069
SMA	34	39,83	95,97	75,2382	13,93851
PT	34	66,09	99,77	87,7797	7,64283
Valid N (listwise)	34				

Dengan menggunakan rumus persamaan (1), diperoleh nilai standardisasi dari masing-masing data adalah sebagai berikut.

- Untuk data ke-1 (Aceh)

$$Z_{SD} = \frac{3{,}33 - 10{,}8276}{6{,}32304} = -1{,}185758749 \approx -1{,}18577$$

$$Z_{SMP} = \frac{24,37 - 47,7409}{16,20069} = -1,442586705 \approx -1,44259$$

$$Z_{SMA} = \frac{52,61 - 75,2382}{13,93851} = -1,623430338 \approx -1,62343$$

$$Z_{PT} = \frac{79,03 - 87,7797}{7,64283} = -1,144824626 \approx -1,14482$$

Perhitungan nilai standardisasi data selanjutnya dilakukan dengan cara yang sama seperti di atas. Hasil nilai standardisasi semua data dapat dilihat dari tampilan "*Data View*" program SPSS berikut.

TABEL 3. Nilai Z-Score SD, SMP, SMA, dan Perguruan Tinggi (PT)

No	Provinsi	Zsd	Zsmp	Zsma	Zpt
1	Aceh	-1,18577	-1,44259	-1,62343	-1,14482
2	Sumatera Utara	-0,32226	-0,17412	-0,04579	0,31406
3	Sumatera Barat	0,06996	0,38018	0,45929	0,10209
4	Riau	-0,09768	0,18265	0,00371	0,12565
5	Jambi	-0,28905	-0,02907	0,22397	0,15967
6	Sumatera Selatan	-0,40133	-0,26239	0,04819	0,72228
7	Bengkulu	-0,29537	-0,12227	-0,10749	-0,49193
8	Lampung	-0,64172	-0,33955	0,32728	-0,56258
9	Kep. Bangka Belitung	0,13955	0,0092	0,56618	0,48023
10	Kepulauan Riau	1,25293	1,30359	0,94212	1,32677

11	DKI Jakarta	2,83287	1,77888	1,27645	1,56883
12	Jawa Barat	0,39891	1,02953	0,97584	0,76677
13	Jawa Tengah	0,49222	1,18446	1,07054	0,71443
14	DI Yogyakarta	2,35683	2,15541	1,48737	1,42621
15	Jawa Timur	0,6741	1,1752	0,80437	0,94341
16	Banten	0,26765	0,46165	0,72904	0,8492
17	Bali	1,8207	1,49988	1,18246	0,96303
18	Nusa Tenggara Barat	-1,03394	-0,93458	-0,61185	-1,22856
19	Nusa Tenggara Timur	-1,25061	-1,72838	-1,69589	-0,56258
20	Kalimantan Barat	-0,43771	-0,50744	-0,53795	-0,17267
21	Kalimantan Tengah	0,14271	0,02464	0,06828	-1,4837
22	Kalimantan Selatan	0,85439	0,87522	0,65945	0,37555
23	Kalimantan Timur	1,01729	0,72461	0,74913	0,80079
24	Kalimantan Utara	0,08736	-0,10375	0,7355	0,74191
25	Sulawesi Utara	0,48273	0,16105	0,14074	0,32714
26	Sulawesi Tengah	-0,75876	-0,49016	-0,23161	-0,48146
27	Sulawesi Selatan	0,00037	0,2092	0,41409	0,6477
28	Sulawesi Tenggara	-0,84574	-0,80125	-0,27465	0,08509
29	Gorontalo	-0,50571	-0,11363	0,16155	0,33107
30	Sulawesi Barat	-1,14306	-1,30926	-0,85291	-0,85697
31	Maluku	-0,6781	-0,62781	-1,31852	-1,17492
32	Maluku Utara	-1,21107	-1,59813	-2,54032	-2,83791
33	Papua Barat	-0,69233	-0,87224	-0,94617	-1,55305
34	Papua	-1,10036	-1,69875	-2,23899	-1,22071

Untuk proses analisis *cluster* selanjutnya, data yang akan digunakan adalah data nilai z-*score* pada tabel di atas.

Menentukan Ukuran Kemiripan

Ukuran kemiripan yang akan digunakan dalam proses *clustering* adalah ukuran jarak. Semakin kecil jarak antara dua objek, maka semakin tinggi tingkat kemiripannya. Dalam penelitian ini mengukur jarak antar dua objek dihitung dengan menggunakan jarak *euclidean* pada persamaan (2). Jarak *euclidean* merupakan jarak suatu garis lurus antara dua titik yang menghubungkan antar objek berupa akar jumlah kuadrat perbedaan nilai untuk tiap variabel.

Proses *Clustering*

Pada penelitian ini, metode *clustering* yang akan digunakan adalah metode *K-means clustering*. Metode ini dimulai dengan menentukan jumlah *cluster* yang ingin dibentuk. Setelah jumlah *cluster* diketahui, barulah proses *cluster*ing dilakukan. Proses *clustering* dengan metode *k-means* pada penelitian ini dilakukan dengan bantuan program SPSS 25. Berikut merupakan tampilan *output* program SPSS.

TABEL 4. *Initial Cluster Centers*

Zscore	***Cluster***		
	1	**2**	**3**
SD	-1,21107	2,83287	,08736
SMP	-1,59813	1,77888	-,10375

SMA	-2,54032	1,27645	,73550
Perguruan Tinggi	-2,83791	1,56883	,74191

Tabel 4 di atas merupakan tampilan awal proses *clustering* sebelum dilakukan iterasi. Tabel tersebut menunjukkan tiga buah *cluster* yang pertama kali terbentuk dengan nilai pusat *cluster* (*centroid*)-nya masing-masing. Kemudian metode *K-Means Clustering* akan melakukan proses iterasi atau realokasi *cluster* yang ada, yang memuat perubahan pada nilai pusat *cluster* (*centroid*). Proses iterasi berhenti jika *centroid* atau anggota tiap *cluster* tidak berubah. Untuk mengetahui berapa kali proses iterasi dilakukan, dapat dilihat dari tampilan *output* program SPSS berikut.

TABEL 5. *Iteration History*

Iteration	***Change in Cluster Center***		
	1	**2**	**3**
1	1,886	,813	,710
2	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 2. The minimum distance between initial centers is 3,473.

Berdasarkan tabel 5 dapat diketahui bahwa proses iterasi untuk mengelompokan 34 objek dilakukan sebanyak 2 kali dan jarak minimum antar pusat *cluster* (*centroid*) awal yaitu 3,473. Adapun tabel pusat *cluster* (*centroid*) akhir dari masing-masing *cluster* setelah iterasi berhenti dapat dilihat dari tampilan *output* program SPSS berikut.

TABEL 6. *Final Cluster Centers*

Z-score	***Cluster***		
	1	**2**	**3**
SD	-1,03691	2,06583	,00145
SMP	-1,27647	1,68444	,15791
SMA	-1,47851	1,22210	,31544
PT	-1,32244	1,32121	,24067

Tabel 6 di atas merupakan rata-rata dari data z-*score* setiap *cluster* yang akan digunakan untuk proses interpretasi terhadap hasil *clustering*. Selanjutnya dapat diketahui jumlah anggota dari setiap *cluster* yang terbentuk melalui tampilan program SPSS berikut.

TABEL 7. *Number of Cases in each Cluster*

Cluster			***Valid***	***Missing***
1	2	3		
8,000	4,000	22,000	34,000	,000

Berdasarkan tabel 7 di atas menunjukan bahwa *cluster* 1 mempunyai 8 anggota, *cluster* 2 mempunyai 4 anggota dan *cluster* 3 mempunyai 22 anggota. Selanjutnya untuk mengetahui provinsi-provinsi mana saja yang masuk ke dalam *cluster* 1, 2 dan 3 dapat dilihat pada tabel 8.

TABEL 8. Tampilan "*Data View*" program SPSS

No	**Provinsi**	**QCL_1**	**QCL_2**	**No**	**Provinsi**	**QCL_1**	**QCL_2**
1	Aceh	1	0,31985	18	Nusa Tenggara Barat	1	0,93638
2	Sumatera Utara	3	0,59237	19	Nusa Tenggara Timur	1	0,93517
3	Sumatera Barat	3	0,30658	20	Kalimantan Barat	3	1,23882
4	Riau	3	0,34762	21	Kalimantan Tengah	3	1,75279

5	Jambi	3	0,36644	22	Kalimantan Selatan	3	1,17413
6	Sumatera Selatan	3	0,80141	23	Kalimantan Timur	3	1,36195
7	Bengkulu	3	0,93924	24	Kalimantan Utara	3	0,70961
8	Lampung	3	1,14301	25	Sulawesi Utara	3	0,51927
9	Kep. Bangka Belitung	3	0,4018	26	Sulawesi Tengah	3	1,34856
10	Kepulauan Riau	2	0,94036	27	Sulawesi Selatan	3	0,42195
11	DKI Jakarta	2	0,81335	28	Sulawesi Tenggara	3	1,41779
12	Jawa Barat	3	1,27696	29	Gorontalo	3	0,60233
13	Jawa Tengah	3	1,44544	30	Sulawesi Barat	1	0,78765
14	DI Yogyakarta	2	0,6228	31	Maluku	1	0,77257
15	Jawa Timur	3	1,49005	32	Maluku Utara	1	1,88624
16	Banten	3	0,83935	33	Papua Barat	1	0,78657
17	Bali	2	0,4733	34	Papua	1	0,87808

Pada tabel di atas kolom QCL_1 menunjukan omor *cluster* dari setiap provinsi dan QCL_2 menunjukan jarak antar objek dengan *centroid*. Maka telah diperoleh hasil *clustering* dengan metode *k-means clustering* sebagai berikut.

TABEL 9. Hasil Akhir *Clustering*

***Cluster* 1**	***Cluster* 2**	***Cluster* 3**
Aceh	Kepulauan Riau	Sumatra Utara
Nusa Tenggara Barat	DKI Jakarta	Sumatra Barat
Nusa Tenggara Timur	DI Yogyakarta	Riau
Sulawesi Barat	Bali	Jambi
Maluku		Sumatra Selatan
Maluku Utara		Bengkulu
Papua Barat		Lampung
Papua		Kep. Bangka Belitung
		Jawa Barat
		Jawa Tengah
		Jawa Timur
		Banten
		Kalimantan Barat
		Kalimantan Tengah
		Kalimantan Timur
		Kalimantan Selatan
		Kalimantan Utara
		Sulawesi Utara
		Sulawesi Tengah
		Sulawesi Selatan
		Sulawesi Tenggara
		Gorontalo
8 anggota	**4 anggota**	**22 anggota**

Berdasarkan tabel 9 di atas dapat diketahui bahwa *Cluster* 1 terdiri dari 8 anggota yaitu Aceh, Nusa Tenggara Barat, Nusa Tenggara Timur, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, dan Papua. Sedangkan *Cluster* 2 terdiri dari 4 anggota yaitu Kepulauan Riau, DKI Jakarta, DI Yogyakarta, dan Bali. Sedangkan *Cluster* 3 terdiri dari 22 anggota yaitu Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep. Bangka Belitung, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, dan Gorontalo.

Interpretasi Hasil *Clustering*

Dalam proses *clustering* setelah *cluster* terbentuk dan diketahui masing-masing anggotanya, tahap berikutnya yaitu melakukan interpretasi terhadap hasil *clustering*. Tahap ini dimulai dengan menganalisis variabel-variabel yang membedakan antara tiga *cluster*. Analisis dilakukan dengan melihat nilai signifikansi (sig.) dan nilai *F* pada tabel ANOVA. Nilai signifikansi (sig.) dan nilai *F* dari masing-masing variabel dapat dilihat dari tampilan *output* program SPSS berikut.

TABEL 10. ANOVA

Z-score	*Cluster*		*Error*			
	Mean Square	*df*	*Mean Square*	*df*	*F*	*Sig.*
SD	12,836	2	,236	31	54,302	,000
SMP	12,466	2	,260	31	47,905	,000
SMA	12,826	2	,237	31	54,103	,000
PT	11,124	2	,347	31	32,070	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Berdarkan tabel 10 di atas dapat diketahui bahwa nilai sig. untuk semua variabel lebih kecil dari 0,05. Hal ini menunjukkan bahwa ada perbedaan yang berarti antara *cluster* 1, *cluster* 2, dan *cluster* 3, yang berhubungan dengan semua variabel tersebut. Kemudian nilai *F* terbesar adalah 54,302 yang dimiliki variabel SD. Hal ini menunjukkan bahwa ada perbedaan yang berarti antara provinsi-provinsi pada ketiga *cluster* yang berhubungan dengan jumlah pengguna internet yang berpendidikan SD.

Selanjutnya akan dianalisis persentase jumlah pengguna internet masing-masing provinsi yang dikelompokan ke dalam masing-masing *cluster* yang dapat dilihat dari tabel nilai pusat *cluster* (*centroid*) akhir seperti di bawah ini.

TABEL 11. *Final Cluster Centers*

Z-score	*Cluster*		
	1	**2**	**3**
SD	-1,03691	2,06583	,00145
SMP	-1,27647	1,68444	,15791
SMA	-1,47851	1,22210	,31544
PT	-1,32244	1,32121	,24067

Berdasarkan tabel 11 di atas, setiap *cluster* dapat didefinisikan sebagai berikut.

1. *Cluster* 1
   Dalam *Cluster* 1 berisi provinsi-provinsi yang mempunyai persentase jumlah pengguna internet berpendidikan SD, SMP, SMA dan Perguruan Tinggi (PT) di bawah rata-rata populasi provinsi yang diteliti. Hal ini dapat dilihat dari nilai negatif (-) pada seluruh variabel *Cluster* 1 yang terdapat pada tabel *Final Cluster Centers*.
2. *Cluster* 2
   Dalam *Cluster* 2 berisi provinsi-provinsi yang mempunyai persentase jumlah pengguna internet berpendidikan SD, SMP, SMA dan Perguruan Tinggi (PT) di atas rata-rata populasi provinsi yang diteliti. Hal ini dapat dilihat dari nilai positif (+) pada seluruh variabel *Cluster* 2 yang terdapat pada tabel *Final Cluster Centers*.
3. *Cluster* 3
   Dalam *Cluster* 3 berisi provinsi-provinsi yang mempunyai persentase jumlah pengguna internet berpendidikan SD, SMP, SMA dan Perguruan Tinggi (PT) di atas rata-rata populasi provinsi yang diteliti. Hal ini dapat dilihat dari nilai positif (+) pada seluruh variabel *Cluster* 3 yang terdapat pada tabel *Final Cluster Centers*.

Untuk mengetahui provinsi-provinsi dengan persentase pengguna internet yang tinggi, sedang atau rendah dapat dilihat dari rata-rata keseluruhan variabel pada tiap *cluster* yang terbentuk. Berikut merupakan rata- rata keseluruhan variabel pada tiap *cluster*.

TABEL 12. Rata-rata Persentase Jumlah Pengguna Internet

	Rata – Rata
***Cluster* 1**	40,91
Cluster 2	72,27
Cluster 3	57,60

Berdasarkan tabel di atas dapat dilihat bahwa dalam mengelompokan provinsi ke dalam masing-masing *cluster* dengan metode *K-Means clustering* dapat ditarik kesimpulan sebagai berikut.

1. *Cluster* 1

   Cluster 1 mempunyai rata-rata persentase jumlah pengguna internet menurut jenjang pendidikan terendah yaitu sebesar 40,91. Artinya pada tahun 2017 provinsi pada *cluster* 1 menjadi provinsi dengan persentase jumlah pengguna internet menurut jenjang pendidikannya rendah. Dapat dikatakan provinsi pada *cluster* 1 merupakan daerah yang memiliki infrastruktur kurang baik untuk mendukung penggunaan internet.
2. *Cluster* 2

   Cluster 2 mempunyai rata-rata persentase jumlah pengguna internet menurut jenjang pendidikan tertinggi yaitu sebesar 72,27. Artinya pada tahun 2017 provinsi pada *cluster* 2 menjadi provinsi dengan persentase jumlah pengguna internet menurut jenjang pendidikannya tinggi. Dapat dikatakan provinsi pada *cluster* 2 merupakan daerah yang memiliki infrastruktur baik untuk mendukung penggunaan internet.
3. *Cluster* 3

   Cluster 3 mempunyai rata-rata persentase jumlah pengguna internet menurut jenjang pendidikan cukup tinggi yaitu sebesar 57,60. Artinya pada tahun 2017 provinsi pada *cluster* 3 menjadi provinsi dengan persentase jumlah pengguna internet menurut jenjang pendidikannya sedang. Dapat dikatakan provinsi pada *cluster* 3 merupakan daerah yang memiliki infrastruktur cukup baik untuk mendukung penggunaan internet.

Dari uraian di atas, dapat diketahui bahwa urutan *cluster* provinsi dengan persentase jumlah pengguna internet menurut jenjang pendidikan tertinggi sampai terendah yaitu *cluster* 2, *cluster* 3, dan *cluster* 1.

KESIMPULAN DAN SARAN

Kesimpulan

Hasil pengelompokan provinsi di Indonesia berdasarkan data persentase siswa yang mengakses internet menurut jenjang pendidikan menggunakan metode *k-means* terbentuk mejadi 3 *cluster*. *Cluster* 1 terdiri dari 8 provinsi, *Cluster* 2 terdiri dari 4 provinsi, dan *Cluster* 3 terdiri dari 22 provinsi.

Interpretasi dari hasil *clustering* menunjukan bahwa provinsi pada *cluster* 1 merupakan daerah yang memiliki infrastruktur kurang baik untuk mendukung penggunaan internet karena mempunyai rata-rata persentase jumlah pengguna internet menurut jenjang pendidikan terendah yaitu sebesar 40,91, sementara provinsi pada *cluster* 2 merupakan daerah yang memiliki infrastruktur baik untuk mendukung penggunaan internet karena mempunyai rata-rata persentase jumlah pengguna internet menurut jenjang pendidikan tertinggi yaitu sebesar 72,27, dan provinsi pada *cluster* 3 merupakan daerah yang memiliki infrastruktur cukup baik untuk mendukung penggunaan internet karena mempunyai rata-rata persentase jumlah pengguna internet menurut jenjang pendidikan cukup tinggi yaitu sebesar 57,60. Oleh karena itu, perlu adanya pemerataan infrastruktur yang mendukung penggunaan internet pada masyarakat Indonesia.

Saran

Perlu dilakukan penelitian lanjutan dengan menggunakan metode *clustering* lainnya, kemudian dilakukan perbandingan yang pada akhirnya dapat diketahui metode terbaik untuk *clustering*. Selanjutnya dibangun perhitungannya dengan menggunakan aplikasi *software* tertentu, sehingga prosesnya menjadi lebih cepat, akurat, dan *applicable*.

UCAPAN TERIMA KASIH

Terima kasih kepada Bapak Dr. Ir. Bagus Sumargo, M.Si. selaku dosen pembimbing yang telah membimbing, mengarahkan, serta memberikan dorongan untuk menyelesaikan penelitian ini. Selain itu, terima kasih kepada seluruh staf PDSI Kementrian Komunikasi dan Informatika atas saran dan ketersediaan data dalam menunjang penelitian ini.

REFERENSI

Abdurrahman, Ginanjar (2016) 'Clustering Data Ujian Tengah Semester (UTS) Data Mining Menggunakan Algoritma K-Means', *Jurnal Sistem & Teknologi Informasi Indonesia*, 1(2).

Arai, K., Barakbah, A. R. (2007) 'Hierarchical K-Means: an algorithm for centroids initialization for K-Means'. *The Faculty of Science and Engineering, Saga University*, 36(1).

Panji, Aditya (2014) 'Hasil Survei Pemakaian Internet Remaja Indonesia' [online]. Available at: https://tekno.kompas.com/read/2014/02/19/1623250/Hasil.Survei.Pemakaian.Internet.Remaja.Indonesia (Accessed: 5 September 2018).

Rivani, Edmira (2010) 'Aplikasi K-means *Cluster* untuk Pengelompokan Provinsi berdasarkan Produksi Padi, Jagung, Kedelai, dan Kacang Hijau Tahun 2009', *Jurnal Mat Stat*, 10(2), pp.122-134.

Sitepu, Robinson, dkk. (2011) 'Analisis *Cluster* terhadap Tingkat Pencemaran Udara pada Sektor Industri di Sumatera Selatan', *Jurnal Penelitian Sains*, 14(3A), 14303.

Zakaria, Muhammad (2018) 'Pengertian Internet Beserta Fungsi dan Manfaat Internet yang Perlu Anda Ketahui' [online]. Available at: https://www.nesabamedia.com/pengertian-fungsi-dan-manfaat-internet-lengkap/ (Accessed: 5 September 2018).