Analisis *Cluster* Non-Hirarki Dengan Menggunakan Metode *K-Modes* pada Mahasiswa Program Studi Statistika Angkatan 2015 FMIPA Universitas Mulawarman

Non-Hierarchical Cluster Analysis Using K-Modes Method on Student of Statistics Major 2015 at the Faculty of Mathematics and Natural Sciences Mulawarman University

Nur Amah[1], Sri Wahyuningsih[2], Fidia Deny Tisna Amijaya[3]
[1]Mahasiswa Program Studi Statistika FMIPA Universitas Mulawarman
[2,3]Dosen Program Studi Statistika FMIPA Universitas Mulawarman
E-mail: nuramah94@yahoo.co.id[1], swahyuningsih@gmail.com[2], fidiadta@gmail.com[3]

Abstract

Cluster analysis is a technique that used to categorize or classify object into clusters or group which is relatively homogeneous. This research aims to know the number of the best cluster used in the selection of Statistics major using K-Modes Cluster, which variable as the best center of cluster & the most optimum, and also comparison of the cluster based on the Davies-Bouldin Index (DBI) which is derived in each cluster are 2 clusters, 3 clusters, and 4 clusters. Steps in this research is descriptive analysis, validity and reliability of questionnaire, determine the number of clusters, compute the dissimiliarity distance, calculate the cluster validation and interpretate the result of the best cluster. Selection of the best cluster use the smallest value comparison. The smallest of the two clusters are 0,599. The center (centroid) of clusters variables which is the best optimum using K-Modes with two clusters are for the first centroid is the first choice of major, SNMPTN, IPK satisfactory, study routines for 4 times a week, and the average length of study is between 60 minutes to 120 minutes per day.; for the second centroid is the first choice of study program, SNMPTN, IPK is very satisfied, study routines for 6 times a week, and the average length of study is less than or equal to 60 minutes per day. The final results showed that the best cluster produced is two clusters where cluster 1 consisted of 37 students and cluster 2 consisted of 8 students.

Keywords : *Cluster analysis, K-Modes cluster, Davies-Bouldin Index (DBI), cluster of validation.*

Pendahuluan

Analisis *cluster* mengklasifikasi objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam *cluster* yang sama. *Cluster-cluster* yang terbentuk memiliki homogenitas internal yang tinggi dan heterogenitas eksternal yang tinggi (Prasetyo, 2014).

Himpunan variabel *cluster* adalah suatu himpunan variabel yang mempresentasikan karakteristik pada objek. Perbedaan analisis *cluster* dengan analisis faktor adalah bahwa analisis *cluster* terfokus pada pengelompokan objek sedangkan analisis faktor terfokus pada kelompok variabel (Prasetyo, 2014).

Metode *K-Modes* merupakan metode *cluster* non-hirarki di mana metode ini dapat berjalan dengan baik untuk himpunan data dengan tipe data berbentuk kategorikal (nominal atau ordinal). Metode ini menggunakan ukuran pencocokan ketidakmiripan sederhana pada fitur data kategorikal, menggunakan modus (nilai yang paling sering muncul) dalam menentukan *cluster* yang menjadi *centroid* (pusat) dan menggunakan metode berbasis frekuensi untuk mencari modus dari sekumpulan nilai yang didapatkan (Prasetyo, 2014).

Penentuan jumlah *cluster* yang diinginkan dalam metode *K-Modes* ini dilakukan di awal sebelum melakukan analisis. Pada penelitian ini, penulis menentukan *cluster* yang akan dibuat adalah 2 *cluster*, 3 *cluster*, dan 4 *cluster* untuk mengetahui jumlah *cluster* yang cocok digunakan dalam penelitian untuk mengelompokkan mahasiswa dengan studi yang dipilihnya berdasarkan pertimbangan pemilihan program studi yaitu program studi statistika. Karena pada prinsipnya, jika jumlah *cluster* berkurang maka homogenitas dalam *cluster* otomatis akan menurun (Supranto, 2010).

Berdasarkan uraian di atas, penulis tertarik untuk mengkaji analisis *cluster* non-hirarki dengan mengambil studi kasus tentang pertimbangan pemilihan program studi Statistika pada mahasiswa Statistika angkatan 2015. Dengan demikian, penulis mengusulkan penelitian berjudul "Analisis *Cluster* Non-Hirarki dengan Menggunakan Metode *K-Modes*".

Uji Validitas

Menurut Sundayana (2015), untuk mengukur tingkat validitas butir kuisioner menggunakan hipotesis sebagai berikut:

H_0 : Butir pertanyaan valid.

H_1 : Butir pertanyaan tidak valid.

Rumus yang dapat digunakan untuk menghitung koefisien korelasi adalah sebagai berikut:

$$r_{hit} = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\left(\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\right)\left(\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}\right)}$$

(1)

dimana,

r_{hit} =nilai korelasi *product moment*
n =jumlah sampel
x =skor item butir soal
y =skor total butir soal

Kriteria penolakan:

H_0 ditolak jika $r_{hit} < r_{\alpha,d\ (n-2)}$ yang berarti butir pertanyaan tidak valid.

Nilai $r_{\alpha,d\ (n-2)}$ didapatkan dengan melihat Tabel Nilai Korelasi *Product Moment*, di mana nilai *α* merupakan taraf signifikansi sebesar 0,05 dan *db* merupakan derajat bebas. Apabila dalam penelitian terdapat pertanyaan yang tidak valid, maka pertanyaan tersebut akan dihilangkan. Selanjutnya, butir pertanyaan yang valid diuji reliabilitasnya.

Uji Reliabilitas

Menurut Sundayana (2015), untuk mengukur pengujian reliabilitas menggunakan hipotesis sebagai berikut:

H_0 : Butir pertanyaan reliabel.
H_1 : Butir pertanyaan tidak reliabel.

Pengujian reliabilitas dalam penelitian ini menggunakan persamaan *Cronbach's Alpha* dengan rumus sebagai berikut:

$$r_{C\ h's.\ ha} = \left(\frac{b}{b-1}\right)\left(1 - \frac{\sum_{i-1}^{n} s_i^2}{s_t^2}\right) \quad (2)$$

dimana,

$r_{C\ h's.\ ha}$ =reliabilitas instrumen
b =banyaknya butir pertanyaan
$\sum_{i-1}^{n} s_i^2$ =jumlahvariansi butir pertanyaan
s_t^2 =variansi total
n =jumlah sampel

Kriteria Penolakan:

H_0 ditolak jika $r_{C\ h's.\ ha} < r_{a,d\ (n-2)}$ dan dapat disimpulkan bahwa butir pertanyaan tidak reliabel.

Analisis *Cluster*

Analisis *cluster* termasuk dalam analisis multivariat metode interdependen. Sebagai alat analisis interdependen maka tujuan analisis *cluster* tidak untuk menghubungkan ataupun membedakan dengan sampel atau variabel lain. Tujuan analisis *cluster* adalah untuk mengidentifikasi sekelompok objek yang mempunyai kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok lainnya. Sehingga objek yang berada dalam kelompok yang sama relatif lebih homogen daripada objek yang berada dalam kelompok yang berbeda. Jumlah kelompok yang dapat diidentifikasi tergantung pada banyak dan variasi data objek (Supranto, 2010).

Metode dalam analisis *cluster* dibagi menjadi dua yaitu, metode hirarki dan non-hirarki. Metode hirarki memulai pengelompokan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang mempunyai kedekatan ke-dua. Dalam metode hirarki *cluster* terdapat dua tipe dasar yaitu *agglomerative* (pemusatan) dan *divisive* (penyebaran). Dalam *agglomerative* ada lima metode yang cukup terkenal yaitu *single linkage*, *complete linkage*, *average lingkage*, W*ard's method*, dan *centroid method*. Dalam *divisive* ada dua metode yaitu *a splinter-average distance method* dan *Automatic Interaction Detection* (AID) (Prasetyo, 2014).

Berbeda dengan metode hirarki, metode non-hirarki dimulai dengan terlebih dahulu menentukan jumlah *cluster* yang diinginkan. Setelah jumlah *cluster* diketahui, kemudian proses *cluster* dilakukan tanpa mengikuti proses hirarki. Dalam metode non-hirarki terdapat tiga metode *clustering* yaitu metode *K-Means*, metode *K-Harmonic Means*, dan metode *K-Modes* (Prasetyo, 2014).

K-Modes Cluster

Analisis *cluster* metode *K-Modes* dapat bekerja dengan baik untuk himpunan data dengan tipe data kategorikal (nominal atau ordinal). Andaikan X dan Y adalah dua data dengan himpunan data berbentuk kategorikal.Ukuran ketidakmiripan diantara X dan Y dapat diukur dengan jumlah ketidakcocokan nilai dari fitur yang berkorespondensi dari dua data.Pengukuran seperti ini sering disebut dengan pencocokan sederhana (*simple matching*) yang diusulkan oleh Kaufman dan Rousseeuw (1990).Rumus yang digunakan seperti pada persamaan Persamaan (3).

$$d(X,Y) = \sum_{j-1}^{r} \in (x_j, y_j) \quad (3)$$

di mana *r* adalah jumlah data, sedangkan $\in$ ( ) adalah nilai pencocokan seperti pada Persamaan (4).

$$\in (x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases} \quad (4)$$

Andaikan X adalah himpunan data yang nilai datanya bertipe kategorikal, $A_1, A_2, \ldots, A_r$,maka modus dari $X = \{X_1, X_2, \ldots, X_n\}$ adalah data $Q = \{q_1, q_2, \ldots, q_r\}$ yang meminimalkan nilai seperti pada Persamaan (5).

$$D(X,Q) = \sum_{i-1}^{n} d(X_i, Q) \quad (5)$$

Untuk Persamaan (5), vektor *Q* merupakan vektor yang bukan bagian dari *X*.

Andaikan $n_{c_{k,j}}$ adalah objek yang dimiliki oleh kategori $c_{k,j}$ ke-*k* pada atribut A_j dan $f_r(A_j = c_k \,|X) = \frac{n_{c_{k,j}}}{n}$ adalah frekuensi relatif kategori $n_{c_{k,j}}$ dalam *X*. Maka fungsi $D(X, Q)$akan minimal jika,

$$f_r(A_j = q_j|X) \geq f_r(A_j = c_{k,j}|X) \qquad (6)$$

untuk$q_j \neq c_{k,j}$ untuk semua $j = 1, \dots, r$.

Fungsi objektif yang digunakan dalam *K-Modes* seperti pada Persamaan (7).

$$J(t) = \sum_{l-1}^{k} \sum_{i-1}^{n} \sum_{j-1}^{r} w_{i,l} \square (x_{i,j}, q_{l,j}) \qquad (7)$$

Di mana □ ( ) adalah nilai pencocokan seperti pada Persamaan (3) antara vektor dengan modus *cluster* yang diikuti, J(*t*) merupakan nilai fungsi objektif pada iterasi ke-*t*, sedangkan $w_{i,l} \square W$ adalah nilai keanggotaan data setiap *cluster*.$w_{i,l} \square W$memiliki nilai [0 1]

$$w_{i,l} = \begin{cases} ji & d(X_1, Q_1) < d(X_1, Q_t), \\ u & l = 1, \dots, k, u \quad t = 1, \dots, k \\ 0, u & t \neq 1 \end{cases} \qquad (8)$$

di mana *k* adalah jumlah *cluster*, sedangkan n adalah jumlah data dalam tiap *cluster* (Prasetyo, 2014).

Menurut Khan (2007), kriteria pemberhentian algoritma *K-Modes* adalah ketika nilai fungsi objektif sekarang yaitu J(t) sama dengan nilai fungsi objektif sebelumnya yaitu J(t-1), maka iterasi dihentikan. Jika tidak sama dengan maka dilanjutkan ke iterasi berikutnya.

Validitas Internal Dengan Menggunakan *Davies-Bouldin Index* (DBI)

Pengukuran *Davies-Bouldin Index* (DBI) diperkenalkan oleh David L. Davies dan Donald W. Bouldin (1979) yang digunakan untuk mengevaluasi *cluster*. Menurut Prasetyo (2014), *Sum of square within cluster* (SSW) sebagai pengukuran kohesi dalam sebuah *cluster* ke-*i* dapat menggunakan perhitungan oleh Persamaan (9).

$$SS_{i} = \frac{1}{m_i} \sum_{j-1}^{m_i} d(x_j, c_i) \qquad (9)$$

di mana m_i adalah jumlah data yang berada dalam *cluster* ke-*i*, sedangkan c_i adalah *centroi dcluster* ke-*i*.

Sementara pengukuran untuk separasi antara dua *cluster*, misalkan *clusteri* dan *j*, digunakan perhitungan *sum of square between cluster* (SSB) dengan mengukur jarak *centroid* c_i dan c_j seperti ada Persamaan (10).

$$S_{i,j} = d(c_i, c_j) \qquad (10)$$

Didefinisikan $R_{i,j}$ adalah ukuran rasio seberapa baik perbandingan antara *cluster* ke-*i* dan *cluster* ke-*j*.$R_{i,j}$ dapat dihitung dengan Persamaan (11).

$$R_{i,j} = \frac{SS_{i} + SS_{j}}{S_{i,j}}$$

Nilai *Davies-Bouldin Index* (DBI) didapatkan dari Persamaan (12).

$$D = \frac{1}{K} \sum_{i-1}^{K} m_{i \neq j}(R_{i,j}) \qquad (12)$$

di mana K adalah jumlah *cluster* yang digunakan (Prasetyo, 2014).

Metodologi Penelitian

Dalam penelitian ini, data yang digunakan adalah data primer yang dikumpulkan dengan cara menyebarkan kuesioner kepada responden penelitian. Populasi dalam penelitian ini adalah seluruh mahasiswa Program Studi Statistika FMIPA UNMUL.Sampel dalam penelitian ini adalah seluruh mahasiswa Program Studi Statistika Angkatan 2015 FMIPA UNMUL yang berjumlah 45 responden.

Uji Validitas

Dilakukan uji validitas dan reliabilitas butir pertanyaan yang telah disebarkan kepada 20 responden dengan pembagian 5 responden mahasiswa Prodi Statistika angkatan 2013, 5 responden mahasiswa angkatan 2014, dan 10 responden mahasiswa angkatan 2015.

Tabel 1. Uji Validitas

Butir Pertanyaan	r_{hitung}	r_{tabel}	Keterangan
1	0,530	0,468	Valid
2	0,585	0,468	Valid
3	0,676	0,468	Valid
4	0,700	0,468	Valid
5	0,720	0,468	Valid

Berdasarkan hasil analisis pada Tabel 1, nilai r_{hit} > $r_{(0,0\ ;1\ )}$ maka diputuskan untuk gagal menolak H_0, sehingga dapat disimpulkan bahwa semua butir pertanyaan keputusan mahasiswa valid.

Uji Reliabilitas

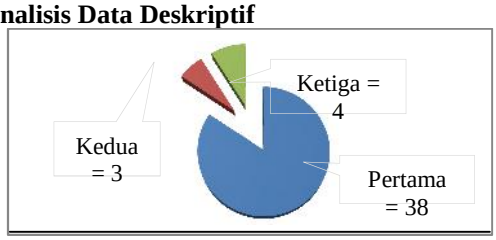
Uji reliabilitas digunakan untuk menguji konsistensi suatu instrumen yang digunakan.

Tabel .2. Uji Reliabilitas

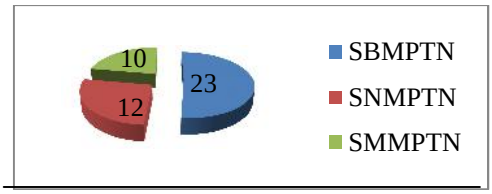
Nilai r_{C} $h's$	Nilai r_t	Keterangan
0,663	0,468	Reliabel

Berdasarkan Tabel 2, dengan menggunakan taraf signifikansi α sebesar 0,05, diperoleh nilai $r_{C\quad n's.\quad na)} > r_{0,0\ ;1}$ yaitu 0,663 > 0,468 maka diputuskan gagal menolak H_0, sehingga dapat disimpulkan bahwa butir pertanyaan keputusan mahasiswa reliabel.

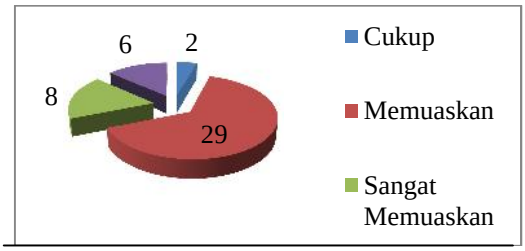
Analisis Data Deskriptif



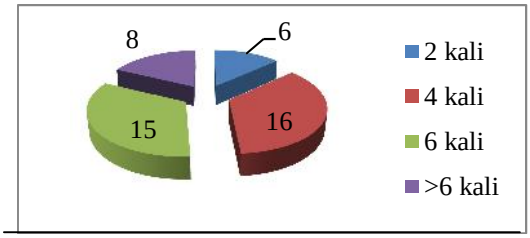
Gambar 1. Karakteristik responden berdasarkan pilihan Program Studi



Gambar 2. Karakteristik responden berdasarkan jalur masuk



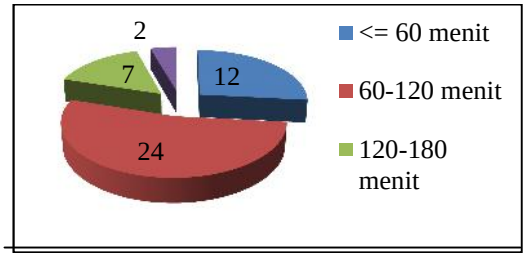
Gambar 3. Karakteristik responden berdasarkan IPK



Gambar 4. Karakteristik responden berdasarkan rutinitas belajar

Analisis K-Modes 2 Cluster
Iterasi Pertama

Dipilih 2 data sebagai modus awal secara acak, penulis memilih data ke-7 dan ke-8.Untuk data ke-7 digunakan sebagai *centroid* 1 dan data ke-8 digunakan sebagai *centroid* 2.$x_i$merupakan nilai dari setiap jawaban responden dan c_i merupakan *modes* awal yang telah dipilih secara acak oleh peneliti.



Gambar 5. Karakteristik responden berdasarkan rata-rata lama belajar

Tabel 3.*Centroid* Awal Iterasi Pertama

Cent-roid	Pilih-an	Jalur	IPK	Rutin-tas	Lama Belajar
1	3	3	2	2	3
2	3	2	3	4	1

Dilakukan iterasi pertama yaitu menghitung ketidakmiripan setiap data ke *centroid* (modus) awal menggunakan Persamaan (2.5) dan Persamaan (2.6).Berikut adalah perhitungan jarak ke setiap *centroid*.

- Data ke-1

$$d(x_1, c_1) = \square(x_1, c_1) + \square(x_1, c_1) + \square(x_1, c_1) + \square(x_1, c_1) + \square(x_1, c_1)$$
$$= \square(1,3) + \square(1,3) + \square(3,2) + \square(3,2) + \square(2,3)$$
$$= 1 + 1 + 1 + 1 + 1 = 5$$
$$d(x_1, c_2) = \square(x_1, c_2) + \square(x_1, c_2) + \square(x_1, c_2) + \square(x_1, c_2) + \square(x_1, c_2)$$
$$= \square(1,3) + \square(1,2) + \square(3,3) + \square(3,4) + \square(2,1)$$
$$= 1 + 1 + 0 + 1 + 1 = 4$$

Perhitungan yang sama dilakukan hingga mencapai data ke *i* = 45. Perhitungan jarak data ke *centroid* dapat dilihat pada Tabel 4. Penentuan letak *cluster* didasarkan pada jarak terdekat anggota data ke-i pada *centroid*. Sebagai contoh, pada data ke-1 jarak data ke *centroid* 1 dan *centroid* 2 masing-masing adalah 5 dan 4 di mana nilai 4 adalah nilai terkecil atau jarak terdekat dari data ke *centroid* 2, maka data ke-1 masuk ke dalam *cluster* 2. Pada data ke-43 jarak data ke *centroid* 1 dan *centroid* 2 masing-masing adalah 4 dan 5 di mana nilai 4 adalah nilai terkecil atau jarak terdekat dari data ke *centroid* 1, maka data ke-43 masuk ke dalam *cluster* 1.

Hasil *cluster* pertama iterasi pertama terdiri dari 31 anggota dapat dilihat pada Tabel 4. Berdasarkan Tabel 4 tersebut dapat dicari modus setiap variabel.Modus didapatkan dengan melihat data yang paling sering muncul pada masing-masing variabel. Apabila terdapat modus dengan jumlah yang sama pada suatu variabel maka diambil salah satu nilai yang digunakan sebagai modus dalam variabel tersebut.Hasil *cluster*

pertama iterasi kedua yang terdiri dari 14 anggota dan didapatkan nilai modus pada *cluster* 2 adalah pilihan program studi pertama, jalur masuk SNMPTN, IPK sangat memuaskan, rutinitas belajar sebanyak 6 kali dalam seminggu, dan rata-rata lama belajar adalah kurang dari sama dengan 60 menit per hari.

Tabel 4. Jarak Data ke *Centroid* Iterasi Pertama

Data ke-	Jarak ke *Centroid*		Terdekat	*Cluster* yang Diikuti
	1	2		
1	5	4	4	2
2	4	3	3	2
3	2	4	2	2
□	□	□	□	□
43	4	5	4	1
44	5	3	3	2
45	4	4	4	1

Modus yang telah terpilih merupakan *centroid* baru yang didapatkan dari kedua *cluster* tersebut. Selanjutnya, modus pada *cluster* 1 merupakan *centroid* 1 dan modus pada *cluster*2 merupakan *centroid* 2.

Nilai fungsi objektif antar setiap data dengan *centroid* yang baru dari masing-masing *cluster* yang diikuti dapat dihitung dengan menggunakan Persamaan (5) untuk mengetahui apakah *cluster* sudah mencapai batas optimal. $X_i$merupakan nilai dari masing-masing data (*X*) yang masuk ke dalam masing-masing *cluster* setelah menghitung nilai ketidakmiripan, sedangkan q_i merupakan nilai modus dari masing-masing data (*X*) pada *cluster* yang telah dihasilkan.

- Untuk *cluster* 1
  Nilai fungsi objektif dapat dihitung berdasarkan persamaan (5).

$$D(X_3, q_1) = \square(X_3, q_1) + \square(X_3, q_1) + \square(X_3, q_1) + \square(X_3, q_1) + \square(X_3, q_1)$$
$$= \square(2,1) + \square(3,1) + \square(2,2) + \square(2,2) + \square(1,2)$$
$$= 1 + 1 + 0 + 0 + 1 = 3$$

$$D(X_5, q_1) = \square(X_5, q_1) + \square(X_5, q_1) + \square(X_5, q_1) + \square(X_5, q_1) + \square(X_5, q_1)$$
$$= \square(1,1) + \square(2,1) + \square(2,2) + \square(3,2) + \square(2,2)$$
$$= 0 + 1 + 0 + 1 + 0 = 2$$

Perhitungan yang sama dilakukan hingga 31 data pada *cluster* 1dan 14 data pada *cluster* 2.
Dengan demikian, diperoleh nilai fungsi objektif pada iterasi pertama adalah sebagai berikut:

$$J(t) = \sum_{l-1}^{k}\sum_{i-1}^{n}\sum_{j-1}^{r} w_{i,l} \square (x_{i,j}, q_{l,j})$$
$$= \sum_{l-1}^{2}\sum_{i-1}^{4}\sum_{j-1}^{5} w_{i,l} \square (x_{i,j}, q_{l,j})$$
$$J(1) = w_{1,1} \square (x_{1,1}, q_{1,1}) + w_{1,1} \square (x_{1,2}, q_{1,2}) + \cdots + w_{2,5} \square (x_{4\ ,5}, q_{2,5}) + w_{2,5} \square (x_{4\ ,5}, q_{2,5})$$
$$= 1(2) + 1(4) + \cdots + 1(3) + 1(2)$$
$$= 83$$
$$\text{Perubahan Fungsi Objektif} = 100 - 83 = 17$$

Didapatkan perubahan nilai fungsi objektif sebesar 17, karena perubahan fungsi objektif belum mencapai kondisi optimal yaitu fungsi objektif sekarang tidak sama dengan fungsi objektif sebelumnya, maka proses *cluster* dilanjutkan ke iterasi berikutnya.
Perhitungan yang sama dilakukan untuk iterasi selanjutnya hingga didapatkan nilai perubahan fungsi objektif sebesar 0.

Tabel 5. Hasil *K-Modes* 2 *Cluster*

Iterasi	Perubahan Fungsi Objektif
Kedua	1
Ketiga	0

Anggota kelompok yang masuk dalam *cluster* 1 terdapat sebanyak 37 data yaitu data ke 1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 13, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 28, 29, 30, 31, 32, 33, 35, 36, 37, 39, 40, 41, 42, 43, 44, dan 45 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SNMPTN, IPK memuaskan, rutinitas belajar sebanyak 4 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah antara 60 menit sampai dengan 120 menit. Dan anggota kelompok yang masuk dalam *cluster* 2 terdapat sebanyak 8 data yaitu data ke 4, 8, 14, 17, 26, 27, 34, dan 38 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SNMPTN, IPK sangat memuaskan, rutinitas belajar sebanyak 6 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah kurang dari sama dengan 60 menit.

Analisis K-Modes 3 *Cluster*

Dipilih 3 data sebagai *modes* awal secara acak, penulis memilih data ke-7, ke-8, dan ke-33.Untuk data ke-7 digunakan sebagai *centroid* 1, data ke-8 digunakan sebagai *centroid* 2, dan data ke-33 digunakan sebagai *centroid* 3.

Tabel 6. Hasil *K-Modes* 3*Cluster*

Iterasi	Perubahan Fungsi Objektif
Pertama	25
Kedua	3
Ketiga	0

Anggota kelompok yang masuk dalam *cluster* 1 terdapat sebanyak 14 data yaitu data ke 3, 6, 7, 10, 11, 12, 13, 21, 23, 24, 35, 37, 41, dan 42 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SMMPTN, IPK memuaskan, rutinitas belajar sebanyak 4 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah antara 120 menit sampai dengan 180 menit. Anggota kelompok yang masuk dalam *cluster* 2 terdapat sebanyak 12 data yaitu data ke 2, 4, 8, 19, 26, 27, 29, 32, 34, 38, 44, dan 45 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SBMPTN, IPK sangat memuaskan, rutinitas belajar sebanyak 4 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah kurang dari sama dengan 60 menit. Dan anggota kelompok yang masuk dalam *cluster* 3 terdapat sebanyak 19 data yaitu data ke 1, 5, 9, 14, 15, 16, 17, 18, 20, 28, 30, 31, 33, 36, 39, 40, dan 43 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SNMPTN, IPK sangat memuaskan, rutinitas belajar sebanyak 2 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah 60 menit sampai dengan 120 menit.

Analisis K-Modes 4 *Cluster*

Dipilih 4 data sebagai modus awal secara acak, penulis memilih data ke-7, ke-8, ke-33, dan ke-44.Untuk data ke-7 digunakan sebagai *centroid* 1, data ke-8 digunakan sebagai *centroid* 2, data ke-33 digunakan sebagai *centroid* 3, dan data ke-44 digunakan sebagai *centroid* 4.

Tabel 7. Hasil *K-Modes* 4 *Cluster*

Iterasi	Perubahan Fungsi Objektif
Pertama	32
Kedua	3
Ketiga	0

Anggota kelompok yang masuk dalam *cluster* 1 terdapat sebanyak 14 data yaitu data ke 3, 6, 7, 10, 11, 12, 13, 21, 23, 24, 35, 37, 41, dan 41 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SMMPTN, IPK memuaskan, rutinitas belajar sebanyak 4 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah 120 menit sampai dengan 180 menit. Anggota kelompok yang masuk dalam *cluster* 2 terdapat sebanyak 11 data yaitu data ke 4, 8, 19, 26, 27, 29, 30, 32, 34, 38, dan 45 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SBMPTN, IPK memuaskan, rutinitas belajar sebanyak 2 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah kurang dari sama dengan 60 menit. Anggota kelompok yang masuk dalam *cluster* 3 terdapat sebanyak 15 data yaitu data ke 9, 14, 15, 16, 17, 18, 20, 22, 25, 28, 31, 33, 39, 40, dan 43 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SNMPTN, IPK memuaskan, rutinitas belajar sebanyak 4 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah 60 menit sampai dengan 120 menit. Dan anggota kelompok yang masuk dalam *cluster* 4 terdapat sebanyak 5 data yaitu data ke 1, 2, 5, 36, dan 44 dengan kategori pilihan Program Studi Statistika adalah pilihan pertama, jalur masuk studi adalah SBMPTN, IPK sangat memuaskan, rutinitas belajar sebanyak 6 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah 60 menit sampai dengan 120 menit.

Validitas *Cluster*

Nilai DBI 2 *Cluster*

Nilai *Sum of Square Within cluster* (SSW) pada *cluster* 1 didapat dengan menggunakan Persamaan (9) adalah:

$$SS_1 = \frac{1}{m_1}\sum_{j-1}^{m_1} d(x_j, c_1)$$
$$= \frac{1}{37}\sum_{j-1}^{3}(2+3+3+\cdots+1+3+2)$$
$$= \frac{1}{37}(68) = 1{,}838$$

Nilai SSW pada *cluster* 1 sebesar 1,838, yang berarti bahwa kedekatan hubungan data dengan *cluster* 1 sebesar 1,838.

Nilai *Sum of Square Within cluster* (SSW) pada*cluster* 2 didapat dengan menggunakan Persamaan (9) adalah:

$$SS_2 = \frac{1}{m_2}\sum_{j-1}^{m_2} d(x_j, c_2)$$
$$= \frac{1}{8}\sum_{j-1}^{8}(2+2+1+2+2+1+2+2)$$
$$= \frac{1}{8}(14) = 1{,}75$$

Nilai SSW pada *cluster* 2 sebesar 1,75, yang berarti bahwa kedekatan hubungan data dengan *cluster* 2 sebesar 1,75.

Nilai *Sum of Square Between cluster* (SSB) kedua*cluster* didapat dengan menggunakan Persamaan (10) adalah:

$$S_{1,2} = d(c_1, c_2)$$
$$= d\big((1,1),(1,1),(2,3),(2,4),(2,1)\big)$$
$$= 0+0+1+1+1 = 3$$

Nilai SSB antara *cluster* 1 dan *cluster* 2 adalah sebesar 3, yang berarti bahwa nilai keterpisahan antar *cluster* adalah berbeda.

Nilai rasio perbandingan antar *cluster* didapat dengan menggunakan Persamaan (11) adalah:

$$R_{i,j} = \frac{SS_i + SS_j}{S._i}$$

$$R_{1,2} = \frac{SS_1 + SS_2}{S._{1,2}} = \frac{1{,}838 + 1{,}75}{3}$$

$$= 1{,}196$$

Nilai rasio perbandingan antara *cluster* 1 dan *cluster* 2 adalah sebesar 1,196, yang berarti bahwa *cluster* 1 dan *cluster* 2 memiliki ukuran rasio sebesar 119,6%.

Nilai *Davies-Bouldin Index* (DBI) didapat dengan menggunakan Persamaan (12) adalah:

$$D = \frac{1}{2}\sum_{i-1}^{2} m_{i\neq j}(R_{1,2}) = \frac{1}{2}(1{,}196) = 0{,}599$$

Tabel 8. Nilai *Davies-Bouldin Index* (DBI)

Cluster	DBI
2	0,599
3	1,289
4	1,353

Berdasarkan Tabel 8, diperoleh nilai DBI pada *cluster* 2, 3, dan 4 masing-masing adalah sebesar 0,599, 1,289, dan 1,353. *Cluster* yang baik merupakan *cluster* yang mempunyai nilai DBI yang paling kecil, maka banyaknya *cluster* yang cocok pada analisis *cluster* non-hirarki dengan menggunakan metode *K-Modes* berdasarkan pemilihan Program Studi Statistika angkatan 2015 adalah dengan menggunakan 2 *cluster* dengan nilai DBI sebesar 0,599.

Kesimpulan

Cluster terbaik yang dihasilkan adalah dengan menggunakan *K-Modes* 2 *cluster* yang mempunyai nilai validitas *Davies-Bouldin Index* (DBI) yang paling kecil yaitu sebesar 0,599.

Variabel yang menjadi pusat (*centroid*) *cluster* terbaik paling optimal untuk *K-Modes* 2 *cluster* yaitu untuk *centroid* 1 adalah pilihan program studi pertama, jalur masuk studi adalah SNMPTN, IPK memuaskan, rutinitas belajar sebanyak 4 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah antara 60 menit sampai dengan 120 menit dan untuk *centroid* 2 adalah pilihan program studi pertama, jalur masuk studi adalah SNMPTN, IPK sangat memuaskan, rutinitas belajar sebanyak 6 kali dalam seminggu dan rata-rata lama belajar mahasiswa per hari adalah kurang dari sama dengan 60 menit.

Hasil dari pengelompokan anggota *K-Modes* 2 *cluster* sebagai *cluster* terbaik yaitu untuk *cluster* 1 terdapat sebanyak 37 mahasiswa dan untuk *cluster* 2 terdapat sebanyak 8 mahasiswa.

Daftar Pustaka

Khan, S.S. 2007.Computation of Initial Modes for K-Modes Clustering Algorithm using Evidence Accumulation.*Proceeding*: IJCAI (*International Joint Conferences on Artifical Intelligence Journal Division*).

Kuncoro, M. 2003. *Metode Riset untuk Bisnis dan Ekonomi*. Jakarta: Erlangga.

Prasetyo, E. 2010.*Data Mining dan Aplikasi Menggunakan Matlab*. Yogyakarta: Penerbit ANDI.

_____. 2014. *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Penerbit ANDI.

Santoso, S. 2014. *Statistik Multivariat Edisi Revisi Konsep dan Aplikasi dengan SPSS*. Jakarta: Elex Komputindo.

Sudjana. 1996. *Metoda Statistika*. Bandung: Tarsito.

Sugiyono. 2007. *Statistika untuk Penelitian.* Bandung : CV. Alfabeta.

Sundayana, R. 2015. *Statistika Penelitian Pendidikan*. Bandung: CV. Alfabeta.

Supranto, J. 2010. *Analisis Multivariat Arti dan Interpretasi*. Jakarta: Rineka Cipta.