Seleksi Atribut Menggunakan Information Gain Untuk Clustering Penduduk Miskin Dengan Validity Index Xie Beni

Femi Dwi Astuti
Program Studi Teknik Informatika
STMIK AKAKOM
femi@akakom.ac.id

Abstrak - Di wilayah Kecamatan Bantul, seorang warga disebut sebagai keluarga miskin berdasarkan beberapa aspek seperti aspek pangan, sandang, papan, penghasilan, kesehatan, pendidikan, kekayaan, air bersih, listrik maupun jumlah jiwa. Aspek-aspek tersebut akan digunakan sebagai atribut dalam proses clustering. Masing-masing atribut memiliki nilai yang akan diolah. Penelitian ini dikerjakan menggunakan seleksi atribut *information gain* sebelum proses *clustering* untuk melihat atribut mana yang sebenarnya berpengaruh dan tidak, sehingga hanya atribut yang berpengaruh saja yang akan digunakan, metode *Fuzzy C-Means* untuk *clustering* penduduk miskin dan *Xie Beni* untuk menentukan jumlah klaster terbaik. Hasil penelitian menunjukkan penggunaan *information gain* dengan *threshold* 0.0001 untuk *clustering* dengan menghilangkan atribut penghasilan memiliki hasil *cluster* yang sama dengan menggunakan atribut penghasilan. Pengujian terhadap 23, 500, 1000 dan 1313 untuk jumlah *cluster* 2, 3, 4, 5, 6 dan 7 menunjukkan bahwa nilai dari *Xie-Beni Index* terkecil adalah 5 dengan nilai 0,1343, sehingga *cluster* yang paling optimal adalah 5.

Kata Kunci : *Clustering, Information Gain,* Kemiskinan*, Fuzzy C-Means, Xie-Beni*

I. PENDAHULUAN

Masalah utama dalam proses *discovering knowledge* dari data penduduk miskin adalah mengidentifikasi atribut yang tepat untuk proses *clustering*. Di wilayah Kecamatan Bantul, seorang warga disebut sebagai keluarga miskin berdasarkan beberapa aspek seperti aspek pangan, sandang, papan, penghasilan, kesehatan, pendidikan, kekayaan, air bersih, listrik maupun jumlah jiwa. Aspek-aspek tersebut akan digunakan sebagai atribut dalam proses *clustering*. Masing-masing atribut memiliki nilai yang akan diolah. Penduduk miskin di Kecamatan Bantul sebanyak 1313 keluarga. Keluarga-keluarga tersebut memiliki karakteristik yang berbeda-beda. Saat ini, penduduk di Kecamatan Bantul dikatakan sangat miskin apabila jumlah nilai dari 11 atribut tersebut tinggi, dengan cara demikian, pemberian bantuan menjadi tidak tepat sasaran karena belum pasti dari aspek mana keluarga tersebut kekurangan. Atribut-atribut yang banyak dan sebenarnya tidak berpengaruh dapat mengurangi kinerja algoritma *clustering*, sehingga perlu dilakukan seleksi atribut untuk mengetahui atribut yang sebenarnya berpengaruh dan yang tidak berpengaruh.

Penggunaan seleksi atribut memiliki hasil akurasi yang lebih baik jika dibandingkan dengan klasifikasi tanpa melalui seleksi fitur [1]. Seleksi fitur yang menunjukkan akurasi tertinggi yaitu *Information Gain* [2]. ”Indeks *Xie Beni* digunakan untuk pengelompokkan karena memiliki ketepatan dan kehandalan yang tinggi untuk digunakan sebagai kriteria dalam menentukan jumlah kelompok yang optimum” [3]. Pemilihan *cluster optimum* pada hasil *clustering* untuk kasus pengelompokkan kabupaten/kota diprovinsi Jawa Tengah berdasarkan variabel pembentuk indeks pembangunan manusia pernah diteliti oleh Purnamasai [4].

Upaya-upaya untuk membantu program pengentasan kemiskinan di daerah Bantul pernah dilakukan melalui beberapa penelitian. Penentuan indikator yang berperan pada identifikasi kemiskinan menggunakan *data mining* pernah dilakukan dan hasilnya menunjukkan 8 indikator dari 11 indikator yang berperan [5]. Penelitian serupa dilakukan untuk klasifikasi keluarga miskin menggunakan *Analytical Hierarchy Process* (AHP) [6].

Teknik seleksi atribut dilakukan untuk mengurangi fitur yang tidak relevan dan mengurangi dimensi atribut pada data. Tujuan penelitan ini adalah untuk mengidentifikasi atribut yang paling berpengaruh maupun yang tidak berpengaruh terhadap proses *clustering* penduduk miskin dengan mengimplementasikan teknik seleksi atribut *Information Gain*. Analisis *information gain attribut evaluation* pernah dilakukan untuk klasifikasi serangan intrusi yang hasilnya menyatakan *information gain* tidak mampu secara signfikan meningkatkan performansi akurasi, deteksi dan *false positive* algoritma *Naive Bayes* [7]. Selain bertujuan untuk menyeleksi fitur/atribut, tujuan lain dari penelitian ini yaitu mencari jumlah *cluster* paling optimal menggunakan validasi index *Xie Beni*. Penelitian ini diterapkan pada algoritma *clustering Fuzzy C-Means*. Pengujian akan dilakukan untuk membandingkan hasil sebelum seleksi atribut dan sesudah seleksi atribut.

Keakuratan hasil cluster perlu diperhatikan sehingga hasil pengelompokkan menjadilebih optimal. Pada penelitian

ini, evaluasi tingkat keakuratan klaster menggunakan Algoritma *Xie-Beni* Index. Algoritma ini merupakan algoritma evaluasi keakuratan klaster yang paling banyak digunakan. Algoritma ini melibatkan kedua matriks U dan kumpulan data.

II. METODOLOGI PENELITIAN

Metode-metode yang digunakan dalam penelitian ini diantaranya adalah *information gain* untuk seleksi atribut, *fuzzy c-means* untuk *clustering* penduduk miskin dan *validity index xie beni* untuk menentukan jumlah *cluster* terbaik.

A. *Information Gain*

Information Gain dari suatu atribut diperoleh dari nilai *entropy* sebelum pemisahan dikurangi dengan nilai *entropy* setelah pemisahan. atribut yang tidak relevan akan menurunkan performa *machine learning*. Sedangkan atribut yang redundan akan membuat *machine learning* bekerja lebih lama [8]. Seleksi fitur menggunakan *information gain* dilakukan dengan cara menghitung nilai *gain* setiap fitur. Ada tiga tahapan dalam pemilihan fitur menggunakan *Information Gain* diantaranya adalah sebagai berikut :

1. Hitung nilai *gain* informasi untuk setiap atribut dalam *dataset* asli
2. Buang semua atribut yang tidak memenuhi kriteria yang ditentukan
3. *Dataset* direvisi

Pengukuran atribut ini dipelopori oleh Claude Shannon pada teori informasi, dituliskan dengan persamaan (1) :

$$Info(D) = -\sum_{i=1}^{m} p_i log_2 p_i \quad (1)$$

Keterangan :
D = himpunan kasus
m = jumlah partisi D
p_i = Proporsi dari D_i terhadap D

Dalam hal ini p_i adalah probabilitas sebuah tuple pada D masuk ke kelas C_i dan diestimasi dengan $| C_i ,D |/|D |$. Fungsi diambil berbasis 2 karena informasi dikodekan berbasis bit. Selanjutnya mencari nilai entropy setelah pemisahan dengan persamaan (2) :

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} x\, I(D_j) \quad (2)$$

Keterangan :
D = Himpunan kasus
A = Atribut
v = jumlah partisi atribut A
$|D_j|$ = jumlah kasus pada partisi ke j
$||D||$ = jumlah kasus dalam D
$I(D_j)$ = total entropy dalam partisi

Untuk mencari nilai *Information Gain* atribut A diperoleh dengan persamaan (3):

$$Gain(A) = I(D) - I(A) \quad (3)$$

Keterangan :
$Gain(A)$ = information gain atribut A
$I(D)$ = total entropy
$I(A)$ = entropy A

Dengan penjelasan lain, $Gain(A)$adalah reduksi yang diharapkan di dalam entropi yang disebabkan oleh pengenalan nilai atribut dari A. Atribut yang memiliki nilai *information gain* terbesar dipilih sebagai uji atribut untuk himpunan S. Selanjutnya suatu simpul dibuat dan diberi label dengan label atribut tersebut, dan cabang-cabang dibuat untuk masing-masing nilai dari atribut.

B. *Validity Index Xie Beni*

Sesuai dengan namanya Indeks XB ditemukan oleh Xie dan Beni yang pertama kali dikemukakan pada tahun 1991. Validitas dalam HCM ditentukan oleh banyak kelompok optimum melalui perhitungan Indeks validitas. Proses validasi *index* dengan Xie Beni membandingkan rata-rata pusat *cluster* akhir dengan data validasi yang diperoleh dari data penduduk miskin, Rumus proses validasi menggunakan Xie-Beni *index* dapat dilihat pada persamaan 4.

$$\mathrm{XB}(c) = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}\left(\left[\sum_{a=1}^{p}(X_{ka} - V_{ia})^2\right](\mu_{ik})\right)}{n.min_{i,j}\left\|V_i, V_j\right\|^2} \quad (4)$$

Formula ini dapat digunakan untuk metode *hard ataupun fuzzy partition* seperti *K-means cluster* maupun FCM. Kriterianya banyak kelompok optimum diberikan oleh nilai XB yang minimum pada lembah pertama. Dengan c menyatakan banyak *cluster*, μ_{ik} adalah tingkat keanggotaan, $\left[\sum_{a=1}^{p}(X_{ka} - V_{ia})^2\right]$ adalah jarak observasi dengan pusat *cluster*,V_i adalah pusat *cluster*, *n* merupakan banyak objek yang akan dikelompokkan, $min_{i,j}\ ||V_i,V_j||^2$ menyatakan jarak minimum antara pusat *cluster* V_i *dan* V_j. Kriteria banyak *cluster* optimum diberikan oleh indeks XB yang minimum. Indeks XB memiliki ketepatan dan keandalan yang tinggi baik untuk memberikan banyak kelompok optimum pada metode *hard partition* seperti *K-means cluster* maupun pada FCM [9].

C. Kemiskinan

"Kemiskinan adalah suatu keadaan seseorang atau keluarga yang serba kekurangan". Indikator yang digunakan untuk menentukan keluarga miskin di Kabupaten Bantul dapat dilihat pada Tabel 1.

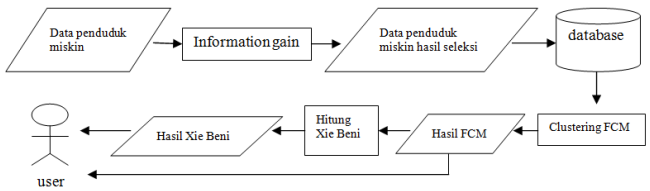
Tabel 1. Indikator Kemiskinan BKKBN

Aspek	Keterangan	Skor
Pangan	Seluruh anggota keluarga tidak mampu makan dengan layak atau senilai Rp. 1.500,- minimal 2 kali dalam sehari	12
Sandang	Lebih dari sebagian anggota keluarga tidak memiliki pakaian pantas pakai minimal 6 stel	9

Aspek	Keterangan	Skor
Papan	Lebih dari 50% Tempat tinggal/ rumah berlantai tanah/ berdinding bambu/ berataprumbia	9
Penghasilan	Jumlah penghasilan yang diterima seluruh anggota keluarga yang berusia 16 tahun keatas < Rp. 993.484	35
Kesehatan	Bila ada anggota keluarga yang sakit tidak mampu berobat ke fasilitas kesehatan dasar	6
Pendidikan	Keluarga tidak mampu menyekolahkan anak yang berumur 7 – 15 tahun	6
Kekayaan 1	Jumlah kekayaan/aset milik keluarga kurang dari Rp.2.500.000,-	5
Kekayaan 2	Tanah bangunan yang ditempati bukan milik sendiri	6
Air Bersih	Tidak menggunakan air bersih untuk keperluan makan, minum & MCK	4
Listrik	Tidak menggunakan listrik untuk keperluan rumah tangga	3
Jumlah Jiwa	Jiwa dalam KK (termasuk kepala keluarga) 5 jiwa atau lebih	5

D. *Gambaran Umum Penelitian*

Gambaran umum penelitian dapat dilihat dari Gambar 1. Berdasarkan Gambar 1 dapat dilihat, proses penelitian dimulai dengan mengumpulkan data penduduk miskin di Kecamatan Bantul. Data yang sudah diperoleh perlu diseleksi terlebih dahulu untuk melihat atribut-atribut yang paling berpengaruh dan yang sebenarnya tidak berpengaruh. proses ini dilakukan dengan menggunakan *information gain*. Selanjutnya *dataset* hasil seleksi atribut disimpan kedalam *database* untuk kemudian dilakukan *clustering* menggunakan metode FCM. hasil *clustering* FCM selanjutnya dievaluasi untuk melihat jumlah *cluster* optimal menggunakan Xie Beni. *User* selain dapat melihat informasi hasil *clustering* FCM juga dapat melihat hasil perhitungan Xie Beni.



Gambar 1. Gambaran Umum Penelitian

III. HASIL DAN PEMBAHASAN

Data yang digunakan adalah data penduduk miskin Kecamatan Bantul yang berjumlah 1313 keluarga miskin dari 5 Desa dan 41 Dukuh. Sebelum dilakukan implementasi sistem, tahap yang dilalui pada penelitian ini yaitu *pre-prosessing* untuk mengetahui atribut yang paling berpengaruh terhadap proses pengelompokkan penduduk miskin. Tujuan seleksi atribut adalah untuk mengidentifikasi beberapa atribut dalam kumpulan data yang sama pentingnya, dan membuang semua atribut lain seperti informasi yang tidak relevan. Langkah untuk seleksi atribut merupakan langkah *pre-processing* yaitu dengan menghilangkan atribut dari data yang tidak relevan sebelum digunakan pada algoritma data mining. Nilai *information gain* terbesar pada atribut suatu data menunjukkan bahwa atribut tersebut adalah atribut yang paling informatif, yang artinya paling relevan terhadap kelas targetnya. Semakin besar nilai *information gain* yang diperoleh pada suatu atribut, maka semakin besar pula pengaruhnya terhadap proses pengelompokkan keluarga miskin. Berdasarkan persamaan yang sudah dijelaskan sebelumnya, total *entropy* yang diperoleh berdasarkan data sebesar 0,3347 sedangkan nilai *entropy* dari masing-masing atribut dapat dilihat pada tabel 2.

Tabel 2. Nilai *Entropy*

No	Atribut	*Entropy*
1	Pangan	0,269810836
2	Sandang	0,325991347
3	Papan	0,313190748
4	Penghasilan	0,334658436
5	Kesehatan	0,304213372
6	Pendidikan	0,33223421
7	Kekayaan 1	0,326143384
8	Kekayaan 2	0,329702732
9	Air bersih	0,306034021
10	Listrik	0,320632292
11	Jumlah Jiwa	0,334154932

Berdasarkan hasil total *entropy* dan *entropy* dari masing-masing atribut, selanjutnya dihitung nilai gain dari masing-masing atribut dengan cara mengurangkan total *entropy* dengan *entropy* dari masing-masing atribut. Hasil perhitungan *information gain* pada setiap atribut untuk penelitian ini dapat dilihat pada tabel 3.

Tabel 3. Nilai *Information Gain*

No	Atribut	*Information Gain*
1	Pangan	0,064906055
2	Sandang	0,008725544
3	Papan	0,021526143
4	Penghasilan	0,00006
5	Kesehatan	0,030503519
6	Pendidikan	0,002482681
7	Kekayaan 1	0,008573507
8	Kekayaan 2	0,005014159
9	Air bersih	0,02868287
10	Listrik	0,014084599
11	Jumlah Jiwa	0,00056

Berdasarkan hasil perhitungan seleksi atribut menggunakan *Information Gain* yang ada pada Tabel 3. dapat dilihat bahwa nilai *Information Gain* yang terendah adalah atribut penghasilan dengan nilai 0,00006 dan nilai *Information Gain* tertinggi adalah atribut pangan dengan nilai *Information Gain* sebesar 0,064906055. Nilai untuk atribut sandang dan kekayaan 1 (kekayaan selain tanah) mempunyai nilai information gain yang hampir sama, nilai atribut sandang sebesar 0,008725544, sedangkan nilai pada atribut kekayaan 1 sebesar 0,008573507. Nilai-nilai *information gain* yang hampir sama juga ada pada atribut air bersih dan atribut papan. Nilai untuk atribut papan sebesar 0,021526143 dan atribut air bersih sebesar 0,02868287. Atribut kesehatan dan listrik memiliki nilai *gain* yang hampir sama dengan atribut papan dan air bersih yaitu sebesar 0,030503519 dan 0,014084599. Nilai atribut pendidikan dan kekayaan 2 berturut-turut sebesar 0,002482681 dan 0,005014159. Atribut jumlah jiwa juga memiliki nilai *information gain* yang rendah sebesar 0,00056.

Dari hasil tersebut dapat disimpulkan bahwa atribut yang paling berpengaruh dalam penentuan kelompok keluarga miskin adalah atribut pangan, dan atribut yang paling tidak berpengaruh adalah atribut penghasilan. Hal ini bertolak belakang dengan kondisi sesungguhnya. Pada kondisi sesungguhnya, data atribut penghasilan memiliki skor nilai yang paling tinggi diantara atribut yang lain yaitu sebesar 35. Secara umum, atribut penghasilan merupakan atribut paling berpengaruh terhadap penentu keluarga miskin, akan tetapi berdasarkan *information gain*, justru atribut penghasilan memiliki nilai *gain* terendah. Setalah dianalisa, hal ini disebabkan karena data penduduk yang dipakai, semua penduduk mempunyai nilai 35 (Jumlah penghasilan yang diterima seluruh anggota keluarga yang berusia 16 tahun keatas < Rp. 993.484). Hasil dari *Information Gain* dapat digunakan sebagai bahan pertimbangan pihak BKKBN maupun pihak pemerintah untuk mengkaji ulang terkait ketentuan besaran penghasilan yang digunakan sebagai dasar penentu keluarga miskin. Setelah diketahui nilai gain dari setiap atributnya langkah selanjutnya adalah menentukan nilai ambang (*Threshold*), dan dalam penelitian ini nilai ambang yang diambil adalah 0,00001 sehingga atribut yang terpilih adalah semua atribut yang ada. Pada penelitian ini akan dicoba dibandingkan proses pengelompokkan penduduk miskin dengan menggunakan 11 atribut dan 10 atribut (dengan menghilangkan atribut penghasilan).

A. Perbandingan hasil *FCM* dengan *IG* dan tanpa *IG*

Pengujian ini akan membandingkan hasil FCM dengan menggunakan seleksi atribut *Information Gain* dan tanpa menggunakan seleksi atribut *Information Gain*. Atribut yang akan digunakan dalam penelitian ini sebanyak 11 atribut (pangan, sandang, papan, penghasilan, kesehatan, pendidikan, kekayaan 1, kekayaan 2, listrik, air bersih dan jumlah jiwa). Hasil penggunaan 11 atribut tersebut akan dibandingkan dengan hasil *clustering* dengan seleksi atribut *information gain*. Nilai *threshold* yang digunakan pada *information gain* yaitu 0,0001. Dengan nilai *threshold* tersebut, atribut penghasilan akan dihilangkan sehingga atribut yang digunakan sebanyak 10 atribut (pangan, sandang, papan, kesehatan, pendidikan, kekayaan 1, kekayaan 2, listrik, air bersih dan jumlah jiwa). Data uji yang digunakan dalam penelitian ini sebanyak 1313 data penduduk.

Berdasarkan tabel 4 dapat dilihat bahwa dengan menggunakan *information gain* dan tanpa menggunakan *information gain* memiliki hasil *cluster* yang sama. Dengan menggunakan *Information Gain* berarti menghilangkan atribut penghasilan. Setelah atribut penghasilan dihilangkan, diperoleh data hasil *cluster*. Hasil *cluster* tanpa atribut penghasilan dibandingkan dengan yang menggunakan atribut penghasilan. Hasil *cluster* tanpa menggunakan atribut penghasilan diperoleh data dalam *cluster* pertama sebanyak 259 keluarga, dengan anggota penduduk nomor {4, 6, 13, 14, 20, 21, 28, 36, 38, 39, 44, 49,…}. *Cluster* kedua sebanyak 297 memiliki anggota penduduk nomor {9, 10, 16, 18, 29, 33, 37, 41, 43, 50, 57,…}. *Cluster* ketiga sebanyak 504, memiliki anggota penduduk nomor {12, 15, 17, 19, 25, 26, 32, 42, 47, 56, 67,…} dan *Cluster* keempat sebanyak 253, memiliki anggota penduduk nomor {1, 2, 3, 5, 7, 8, 11, 22, 23, 24, 27, 30, 31,…}.

Tabel 4. Uji Perbandingan Dengan IG dan Tanpa IG

Cluster	**Dengan IG (menghilangkan atribut penghasilan)**		**Tanpa IG**	
	Jmlh	**Data dalam *cluster***	**Jmlh**	**Data dalam *cluster***
C1	259	{4,6,13,14,20,21,28,36,38,39,44,49,…}	504	{12,15,17,19,25,26,32,42,47,56,67,…}
C2	297	{9,10,16,18,29,33,37,41,43,50,57,…}	297	{9,10,16,18,29,33,37,41,43,50,57,…}
C3	504	{12,15,17,19,25,26,32,42,47,56,67,…}	259	{4,6,13,14,20,21,28,36,38,39,44,49,…}
C4	253	{1,2,3,5,7,8,11,22,23,24,27,30,31,…}	253	{1,2,3,5,7,8,11,22,23,24,27,30,31,…}

Hasil *cluster* menggunakan atribut penghasilan diperoleh data dalam *cluster* pertama sebanyak 504 keluarga, dengan anggota penduduk nomor {12, 15, 17, 19, 25, 26, 32, 42, 47, 56, 67,…}. *Cluster* kedua sebanyak 297 memiliki anggota penduduk nomor {9, 10, 16, 18, 29, 33, 37, 41, 43, 50, 57,…}. *Cluster* ketiga sebanyak 259, memiliki anggota penduduk nomor {4, 6, 13, 14, 20, 21, 28, 36, 38, 39, 44, 49,…} dan *Cluster* keempat sebanyak 253, memiliki anggota penduduk nomor {1, 2, 3, 5, 7, 8, 11, 22, 23, 24, 27, 30, 31,…}.

Berdasarkan hasil tersebut, dapat dilihat bahwa dalam satu *cluster*, memiliki anggota kelompok yang sama. Sebagai contoh, dengan menggunakan atribut penghasilan (tanpa *information gain*) penduduk nomor 12 menjadi satu *cluster* dengan penduduk nomor 15, 17, 19, 25, 26, 32, 42, 47, 56, 67 dan lainnya. Dengan menggunakan *information gain*

(tanpa atribut penghasilan), penduduk nomor 12 juga menjadi satu *cluster* dengan penduduk nomor 15, 17, 19, 25, 26, 32, 42, 47, 56, 67 dan lainnya. Kondisi tersebut juga berlaku untuk *cluster-cluster* yang lain. Dari pembahasan tersebut dapat disimpulkan bahwa sebenarnya atribut penghasilan tidak berpengaruh terhadap proses *clustering* penduduk miskin. Hal ini dapat dijadikan pertimbangan bagi pihak pemerintah Kabupaten Bantul untuk mencermati ulang kriteria penghasilan suatu keluarga.

B. Perhitungan Xie Beni

Nilai *Xie Beni index* digunakan untuk menentukan jumlah *cluster* yang paling optimal. Pada tabel 5 dapat dilihat beberapa jumlah data dan jumlah *cluster* yang digunakan sebagai data uji. Jumlah data yang dipakai adalah 23 (diambil dari data keluarga yang ada di Dukuh Karanggayam), 500, 1000 dan 1313. Jumlah *cluster* yang dicoba pada masing-masing data yaitu 2, 3, 4, 5, 6 dan 7. Berdasarkan hasil validasi *index* dapat dilihat bahwa nilai validasi *index* pada masing-masing *cluster* yang diujikan berbeda.

Tabel 5. Hasil Validasi *Index* Xie Beni

Jumlah Data	Jumlah *Cluster*	Xie Beni
23	2	0.1534
	3	0.1194
	4	0.1042
	5	0.0606
	6	0.1166
	7	0.1329
500	2	0.5895
	3	0.3566
	4	0.1687
	5	0.1569
	6	0.3350
	7	0.4210
1000	2	0.6400
	3	0.3367
	4	0.1683
	5	0.1467
	6	0.3477
	7	0.5675
1313	2	0.6773
	3	0.2385
	4	0.1704
	5	0.1343
	6	0.2509
	7	0.3098

Berdasarkan tabel 5, Nilai *Xie Beni* pada 23 data paling kecil adalah pada saat jumlah clusternya 5 dengan nilai Xie Beni 0.0606, Nilai *Xie Beni* pada 500 data paling kecil adalah pada saat jumlah clusternya 5 dengan nilai *Xie Beni* 0.1569. Nilai *Xie Beni* pada 1000 data paling kecil adalah pada saat jumlah *cluster*-nya 5 dengan nilai *Xie Beni* 0.1467. Nilai *Xie Beni* pada 1313 data paling kecil adalah pada saat jumlah *cluster*-nya 5 dengan nilai *Xie Beni* 0.1343. Berdasarkan data tersebut maka dapat disimpulkan bahwa nilai *Xie Beni* terkecil diperoleh pada jumlah *cluster* yang paling banyak dengan jumlah *cluster* optimal sebanyak 5.

IV. KESIMPULAN

Berdasarkan hasil dan pembahasan yang telah dipaparkan sebelumnya maka dapat diambil kesimpulan sebagai berikut:

1. Penggunaan *information gain* dengan *threshold* 0.0001 untuk *clustering* dengan menghilangkan atribut penghasilan memiliki hasil *cluster* yang sama dengan menggunakan atribut penghasilan.
2. Hasil pengujian terhadap 23, 500, 1000 dan 1313 untuk jumlah *cluster* 2, 3, 4, 5, 6 dan 7 menunjukkan bahwa nilai dari *Xie-Beni Index* terkecil adalah 5 dengan nilai 0,1343, sehingga *cluster* yang paling optimal adalah 5.

REFERENSI

[1] Azhagusundari, B. dan Thanamani, A.S. (2013). Feature Selection Based on Information Gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN:2278-3075, Volume-2, Issue-2, pp. 18-21.

[2] Chandani, V., Wahono, R.S., dan Purwanto. (2015). Komparasi Algoritma Klasifikasi Machine Learning dan Feature Selection pada Analisis Sentimen Review Film. *Journal of Intelligent Systems*. Vol.1, No.1, February 2015, pp. 55-59.

[3] Duo, C., Xue, L., dan Du-Wu, C.. (2007). An Adaptive Cluster Validity Index for the Fuzzy C-Means. *International Journal of Computer Science and Network Security*. Vol. 7, No.2. 146-156.

[4] Purnamasari, S.B., Yasin, H. dan Wuryandari, T. (2014). Pemilihan Cluster optimum pada Fuzzy C-Means (Studi Kasus : Pengelompokan Kabupaten/Kota di provinsi Jawa Trngah Berdasarkan Indikator Indeks Pembangunan Manusia). *Jurnal Gaussian*, volume 3. Nomor 3. Halaman 491-498.

[5] Sela, E.I. (2015). Penentuan Indikator yang Berperan Pada Identifikasi Kemiskinan Menggunakan Data Mining. *Jurnal Riset Daerah Edisi Khusus*, hal 16-32.

[6] Redjeki, S., Guntara, M., dan Anggoro, P. (2014). Perancangan Sistem Identifikasi dan Pemetaan Potensi Kemiskinan untuk Optimalisasi Program Kemiskinan. *Jurnal Sistem Informasi (JSI)*. Vol.6, No.2 Oktober 2014. ISSN : 2085-1588. hlm 731-743.

[7] Essra, A., Rahmadani, Safriadi. (2016). Analisis Information Gain Attribute Evaluation untuk Klasifikasi Serangan Intrusi. *Jurnal ISD* Vol.2 No.2 Juli-Desember. ISSN : 2528-5114. hal 9-14.

[8] Marcelloni, F., (2003). Feature Selection Based on A Modified Fuzzy C-Means Algorithm With Supervision, *Information Sciences*. 151. pp.201-206.

[9] Sen, A., Foster, J. (1997). *On Economic Inequality. Clarendon Paperback*. New York: Oxford University Press.