KOMPUTASI CERDAS

Klastering Data Bahan Makanan Pokok Menggunakan K-Means Clustering

Millatul Ulya[1)]

[1)]*Staf Pengajar Prodi Teknologi Industri Pertanian, Universitas Trunojoyo Madura*
Jl. Raya Telang PO BOX 2 Kamal, Bangkalan, Madura
qumil_2307@yahoo.co.id

Abstrak

Penelitian ini bertujuan untuk mengetahui jumlah klaster yang sesuai untuk data set bahan makanan pokok, dan Membandingkan nilai siluet dan *Sum of Squares Error* (SSE) antara tiga metode *k-means* dalam klastering varietas data set bahan makanan pokok. Hasil penelitian ini menunjukkan bahwa jumlah klaster yang paling sesuai pada data set bahan makanan pokok adalah 3 klaster dengan nilai rata-rata Siluet paling tinggi yakni 0,9089 dan nilai *Sum Square Error* paling kecil sebesar 1,3788e+003.

Kata Kunci **: klastering *k-means*, bahan makanan pokok**

I. PENDAHULUAN

Indonesia merupakan negeri yang kaya akan sumber bahan makanan pokok. Beras adalah makanan pokok yang umum dikonsumsi di Indonesia. Selain beras, masih ada bahan-bahan lain yaitu umbi-umbian dan serealia. Masing-masing bahan makanan pokok memiliki sifat fisikokimia yang berbeda, yakni kadar kalori, kadar protein, kadar lemak, kadar air, kadar abu, kadar vitamin dan mineral. Masing-masing bahan makanan pokok memiliki kemiripan jika dilihat dari beberapa sifat fisiko kimianya. Jika bahan makanan dapat dikelompokkan menjadi beberapa kelompok berdasarkan sifatnya, maka konsumen dapat memilih bahan makanan pokok pengganti yang memiliki sifat yang mirip dengan bahan makanan yang umum dikonsumsi. Bahan makanan pokok nantinya dapat terbagi menjadi beberapa kelompok yang memiliki kemiripan sifat fisikokimia.

Pengelompokan data adalah salah satu teknik dalam data mining. Salah satu teknik penelompokan adalah *klastering*. Tujuan utama dari klastering adalah mengelompokkan sejumlah data/obyek ke dalam klaster-klaster sehingga tiap klaster akan berisi data yang jaraknya dekat atau semirip mungkin (Santosa, 2007). Teknik klastering yang paling umum digunakan adalah k-means clustering. K-means adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi (Agusta, 2007). Dalam teknik ini kita ingin mengelompokkan obyek ke dalam k kelompok atau klaster. Untuk melakukan klastering ini, nilai k harus ditentukan terlebih dahulu.

Penelitian ini bertujuan untuk mengetahui jumlah klaster yang sesuai untuk data bahan makanan pokok dan mengukur performansinya serta membandingkan performansi antara metode klastering kmeans dengan menhhunakan jarak *cityblock*, *euclidean* dan *kernel k-means*. Ukuran performansi yang digunakan untuk data set bahan makanan pokok adalah silhouette value (nilai siluet) dan *Sum of Squares Error* (SSE). Penelitian dilakukan dalam 2 eksperimen yaitu klastering data padi tersebut menjadi 3 dan klaster.

II. METODE

A. *Implementasi pada data set Bahan Makanan Pokok (BMP)*

Untuk mengetahui jumlah klaster yang sesuai pada data set BMP ini, maka dalam penelitian ini akan diimplementasikan pada 2 eksperimen (3 klaster dan 4 klaster) dengan menggunakan k-means dengan jarak *cityblock*, *Euclidean* dan *kernel k-means* kemudian dihitung performansinya untuk mengetahui berapa jumlah klaster yang sesuai untuk data set BMP tersebut.

1. Data set BMP

Data set ini terdiri dari 60 jenis bahan makanan pokok dan memiliki 4 atribut yaitu: kadar kalori (kkal), kadar protein (%), kadar lemak (%), dan kadar air (%). Data set diperoleh dari Rahmawati dan Fatatie (2015).

2. Mengklasterkan data set BMP dengan k-means

Klastering data set BMP dilakukan pada 2 eksperimen, yaitu diklasterkan menjadi: 3 dan 4 klaster. Hal ini dilakukan untuk mengetahui berapa jumlah klaster yang sesuai untuk data set BMP tersebut berdasarkan nilai siluet dan plot siluet yang diperoleh dari hasil klastering.

3. Mengukur Performansi dengan nilai siluet (Ulya, 2010)

Menurut Martinez & Martinez (2005), nilai siluet dapat mengestimasikan jumlah klaster yang sesuai pada data set. Kita notasikan c_i adalah klaster yang berisi i data (obyek). a_i adalah rata-rata *dissimilarity* dari data ke-i ke semua anggota pada klaster yang sama atau $c(i)$. Untuk setiap klaster yang lain, kita notasikan c, maka kita hitung $\bar{d}(i,c)$ mewakili rata-rata *dissimilarity* dari data ke-i ke semua obyek pada klaster c. Kita notasikan b_i sebagai minimum dari rata-rata *dissimilarity* $\bar{d}(i,c)$. Maka, nilai siluet (*silhouette value* atau *silhouette width*) dapat dirumuskan sebagai berikut:

$$sw_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \qquad (1)$$

Kita juga dapat mencari rata-rata nilai siluet dengan merata-rata nilai s_{iK} untuk semua observasi sebagai berikut:

$$\overline{sw} = \frac{1}{n}\sum_{i=1}^{n} sw_i \qquad (2)$$

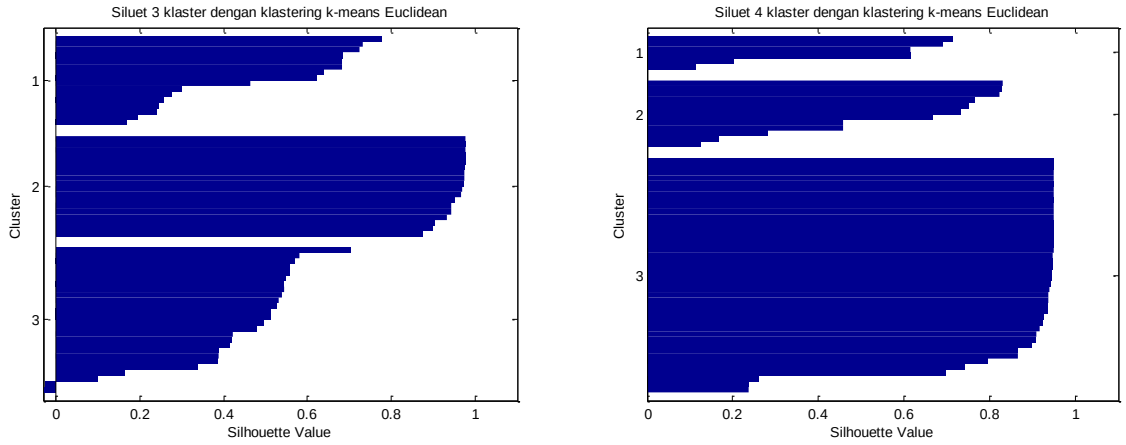
Nilai siluet terletak antara -1 dan 1 ($-1 < sw_i < 1$). Nilai siluet positif yang besar dari sw_i menunjukkan bahwa data/obyek ke-i terklaster dengan baik. Nilai negatif yang besar dari sw_i menunjukkan adanya klastering yang jelek, dan jika nilai dari sw_i mendekati nol mengindikasikan bahwa data/obyek ke-i terletak antara dua klaster. Jika max nilai siluet < 0.25, menunjukkan bahwa tidak ada klaster yang terdefinisi di dalam data tersebut, bahkan mungkin prosedur klastering yang digunakan tidak dapat menemukan klaster-klasternya (Izenman, 2008). Performansi metode klastering juga dapat dilihat dari nilai *Sum Square Error* (SSE) dari klaster yang dihasilkan. Nilai SSE merepresentasikan homegenitas intra klaster, semakin kecil nilai SSE maka semakin homogen data dalam satu klaster. Jadi metode yang terbaik akan memiliki SSE yang paling kecil. Nilai *Sum of Squares Error* sesuai persamaan (3).

$$SSE = \sum (y_{ir} - \bar{c}_r)^2 \qquad (3)$$

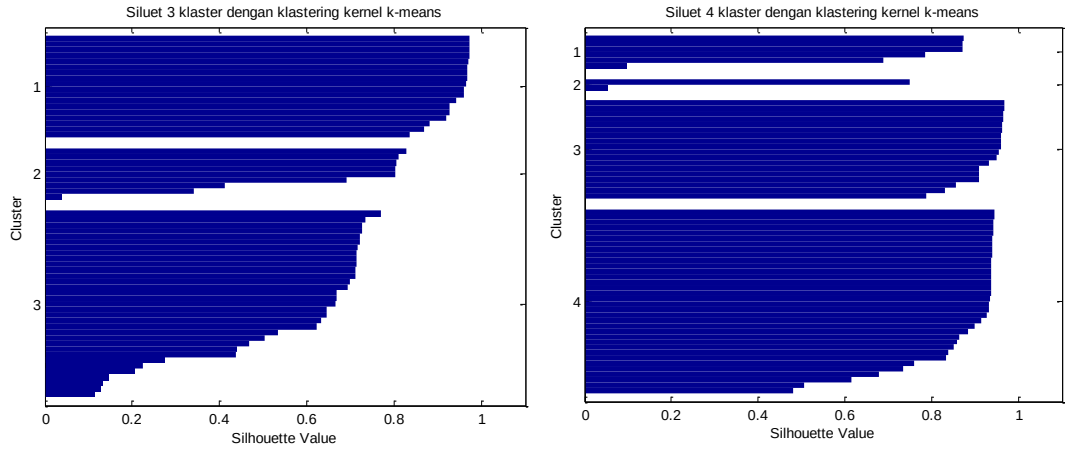
Semua tahapan dilakukan dengan bantuan perangkat lunak Matlab 7.01.

III. HASIL DAN DISKUSI

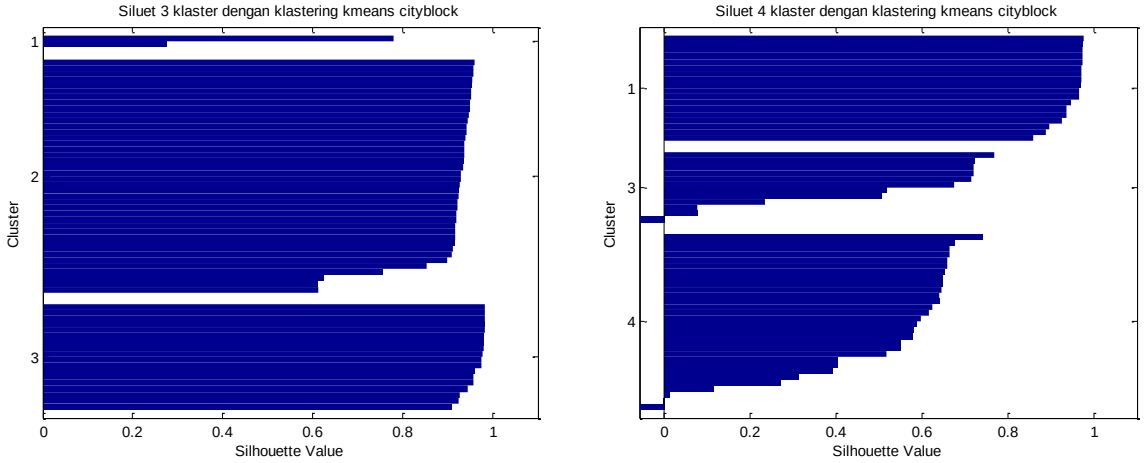
Hasil klastering dengan total terdiri dari 6 eksperimen yaitu kmeans dengan menggunakan rumus jarak euclidean, cityblock dan kernel kmeans masing-masing 3 dan 4 klaster dapat dilihat pada Gambar di bawah ini, yang merupakan gambar nilai siluet dari 6 eksperimen yang telah dilakukan.



Gambar 1. Siluet klastering kmeans dengan jarak Euclidean pada 3 dan 4 klaster pada data set Bahan Makanan Pokok



Gambar 2. Siluet klastering kmeans dengan jarak Cityblock pada 3 dan 4 klaster pada data set Bahan Makanan Pokok



Gambar 3. Siluet klastering kernel kmeans pada 3 dan 4 klaster pada data set Bahan Makanan Pokok

Tabel I.

REKAPITULASI NILAI RATA-RATA SILUET DAN SSE PADA KLASTERING DATA SET BAHAN MAKANAN POKOK

Metode Klastering	Eksperimen	Rata-rata nilai siluet	SSE
Kmeans dengan jarak Euclidean	3 klaster	0,7342	1,6184e+003
	4 klaster	0,8083	2,2333e+003
Kmeans dengan jarak city block	**3 klaster**	**0,9089**	**1,3788e+003**
	4 klaster	0,6288	1,6971e+003
Kernel kmeans	3 klaster	0,6783	8,1613e+008
	4 klaster	0,6368	NaN

Berdasarkan Tabel 1 di atas, dapat dilihat kolom rata-rata nilai siluet untuk menentukan berapa jumlah klaster yang sesuai pada data set BMP. Dari beberapa ekperimen di atas memberikan rata-rata nilai siluet paling tinggi, yaitu 0.9089 dan SSEnya juga paling rendah yaitu 1,3788e+003. Nilai SSE yang semakin kecil menunjukkan bahwa homogenitas intra klaster semakin tinggi. Sehingga metode klastering terbaik adalah yang memiliki SSE terkecil. Plot siluet menunjukkan nilai positif pada semua klaster (lihat Gambar 3). Sehingga, jumlah klaster yang paling sesuai untuk data set Bahan makanan pokok adalah 3 klaster dengan menggunakan metode klastering kmeans dengan rumus jarak cityblock karena memberikan nilai performansi yang paling baik.

Klastering kernel kmeans menunjukkan rata-rata siluet yang paling rendah dan nilai SSE paling tinggi. Di dalam kernel k-means diharapkan data bisa dipisahkan dengan lebih baik karena data yang overlap / non-linier bisa menjadi linier di ruang dimensi baru. Dalam hal ini, pencarian jarak antara tiap titik terhadap pusat klaster dilakukan di *feature space* (Santosa, 2007). Dalam penelitian ini akan dilakukan penerapan metode kernel pada teknik klastering k-means dengan menggunakan fungsi kernel polynomial dengan menggunakan nilai pangkat/degree=2. Namun ternyata hasilnya memberikan nilai *Sum Square Error* yang paling tinggi. Hal ini menunjukkan bahwa data set bahan makanan

pokok bersifat linier atau tidak overlap, sehingga tidak perlu metode kernel dalam pengklasterannya.

Hasil klastering data set bahan makanan pokok menjadi 3 klaster dapat dilihat pada Tabel 2. Pada Tabel tersebut dapat dilihat tentang perbedaan bahan makanan pokok dari masing-masing klaster berdasarkan empat sifat fisikokimianya.

Tabel II.
HASIL KLASTERING DATA SET BAHAN MAKANAN POKOK DENGAN *K-MEANS* CITYBLOCK 3 KLASTER

Klaster	Jumlah Anggota	Anggota Klaster
1	2	Biskuit dan kacang tanah
2	40	Barley, Beras/Nasi, Beras Ketan, Beras ketan hitam, Beras merah, Beras Paboilet, Bihun, Catel, Gandum, Gaplek, Garut, Havermout, Jagung, Jagung giling kuning, jagug giling putih, Jali, Juwawut, Kacang Gedde Kacang Hijau, Kacang Kedelai, Kacang Tunggak, Katul Beras, Katul Jagung, Makaroni, Mi Kering, Oats, Roti warna coklat, roti putih, sagu, Singkong, Sorghum, Tepung Beras, Tepung Gaplek, Tepung Garut, Tepung Kentang, Tepung Sukun Tua, Tepung Tapioka, Tepung Terigu, Tepung Uwi, dan Vermicelli
3	18	Gadung, Ganyong, Gembili, Jagung bubur, Kentang, kentang hitam, mi basah, Sente, Sukun muda, Sukun tua, Suweg, Talas, Tepung sagu, Ubi jalar kuning, ubi jalar merah, ubi jalar putih, Uwi, dan Waluh.

Klaster 1 hanya berisi 2 bahan makan pokok yaitu biskuit dan kacang tanah yang keduanya memiliki kadar kalori paling tingi diantara bahan makanan pokok yang lain.. Sedangkan klaster 2 merupakan klaster yang berisi bahan makanan pokok dengan kadar kalori sedang, termasuk beras. Jika masyarakat Indonesia ingin mengganti beras dengan makanan pokok yang lain, maka dapat memilih bahan makanan pokok yang ada di Klaster 2 karena nilai kalorinya hampir sama. Klaster 3 secara umum berisi umbi-umbian dan bahan makanan pokok lain yang kalorinya rendah.

IV. KESIMPULAN

Klastering data set bahan makanan pokok menghasilkan 3 klaster menggunakan metode kmeans dengan jarak city block memberikan silhouette value yang terbaik yaitu dengan nilai rata-rata siluet 0.9089 dan sum of squares error yang terkecil yaitu 1,3788e+003 dibandingkan metode klastering yang lain. Klaster 1 berisi bahan makanan pokok dengan kalori tinggi, Klaster 2 dengan kalori sedang dan Klaster 3 dengan kalori rendah.

V. DAFTAR PUSTAKA

[1] Agusta, Y. (2007), K-means, Penerapan, Permasalahan dan Metode Terkait, Jurnal Sistem dan Informatika, Vol, 3 (Pebruari 2007), 47-60.

[2] Martinez, A.L. dan Martinez, A.R. (2005), Exploratory Data Analysis with MATLAB, CRC Press Company, USA.

[3] Rahmawati, A, dan Fatatie, Y. (2015). Database Kandungan Gizi pada Bahan Makanan Pokok. http://azaima.tripod.com diakses tanggal 21 November 2015.

[4] Ruskandar, A. (2009), Varietas Ciherang Makin Mendominasi, Warta Penelitian dan Pengembangan Pertanian, Vol.31, No.6.

[5] Santosa, B. (2007), Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis, Graha Ilmu, Jakarta.