JURNAL MATEMATIKA "MANTIK"
Edisi: Oktober 2016. Vol. 02 No. 01
ISSN: 2527-3159 **E-ISSN: 2527-3167**

PERBANDINGAN PENGKLUSTERAN DATA *IRIS* MENGGUNAKAN METODE *K-MEANS* DAN *FUZZY C-MEANS*

Fitria Febrianti[1], Moh. Hafiyusholeh[2], Ahmad Hanif Asyhar[3]
Fakultas Sains dan Teknologi Universitas Islam Negeri Sunan Ampel Surabaya
E-mail: fitriafebrianti09@gmail.com[1], hafiyusholeh@uinsby.ac.id[2], hanif@uinsby.ac.id[3]

Abstrak

Indonesia dengan kekayaan alam yang melimpah, tentu memiliki banyak tanaman yang tak terhitung banyaknya. Untuk mengklaster tanaman menjadi beberapa kelompok yang berbeda dapat menggunakan beberapa metode. Salah satunya metodenya adalah K-Means dan Fuzzy C-Means. Akan tetapi, dua metode ini memiliki perbedaan. Tidak hanya dari segi algoritma, akan tetapi dari segi perhitungan nilai *root mean square error (RMSE)*-nya juga berbeda. Untuk menghitung nilai *RMSE* ada dua indikator yang diperlukan, yaitu data *training* dan data *checking*. Dari pembahasan, metode Fuzzy C-Means memiliki tingkat *RMSE* yang lebih kecil dibandingkan metode K-Means yaitu pada 80 data *training* dan 70 data *checking* dengan nilai *RMSE* 2,2122E-14. Hal ini menunjukkan bahwa metode Fuzzy C-means memiliki tingkat ketepatan yang lebih tinggi dibandingkan dengan metode K-Means.

Kata kunci: *data iris, logika fuzzy, fuzzy c-means, data mining, k-means*

Abstract

Indonesia with abundant natural resources, certainly have a lot of plants are innumerable. To clasify the plants into different clusters can use several methods. Methods used are K-Means and Fuzzy C-Means. However, this methods have difference. Not only in terms of algorithms, but in terms of value calculation on the root mean square error (RMSE) also different. To calculate the value of RMSE there are two indicators are required, namelt the training data and the checking data. Of discussion, the Fuzzy C-Means method has RMSE values smaller than the K-Means method, namely on 80 training data and 70 checking data with RMSE value 2,2122E-14. This indicates that the Fuzzy C-Means method has a higher level of accuracy than the K-Means method.

Kata kunci: *iris data, fuzzy logic, fuzzy c-means, mining data, k-means*

1. Pendahuluan

Indonesia merupakan negara yang kaya akan sumber daya alamnya, oleh karena itu Indonesia memiliki begitu banyak ragam tumbuhan dan bunga yang tersebar diwilayah Indonesia. Dari sekian banyak tumbuhan di Indonesia, hanya 20% yang sudah teridentifikasi [1]. Pada umumnya, beberapa tanaman yang belum diidentifikasi diklaster atau dikelompokkan menjadi beberapa kelompok. Pengklasteran atau pengelompokkan adalah pengelompokan objek atau kasus menjadi kelompok-kelompok yang lebih kecil, dimana setiap kelompok berisi objek atau kasus yang mirip satu sama lain [2]. Terdapat pengklasteran beberapa jenis bunga berdasarkan lebar

ISSN: 2527-3159 E-ISSN: 2527-3167

mahkota, panjang mahkota, lebar kelopak dan panjang kelopak yang sering disebut dengan data iris.

Data iris merupakan data dari 150 bunga yang diidentifikasi berdasarkan panjang mahkota, lebar mahkota, panjang kelopak dan lebar kelopak [3]. Dari 150 data tersebut pada umumnya peneliti-peneliti sebelumnya mengelompokkan menjadi tiga kelompok bunga, yaitu *iris setosa, iris virginica* dan *iris versi color* [3][4][5]. Untuk menguji metode pengklasteran banyak peneliti-peneliti sebelumnya yang menggunakan data iris, karena data iris merupakan data sederhana yang mudah didapat. Ada beberapa metode yang dapat digunakan untuk mengelompokkan data menjadi beberapa kelompok data, diantaranya adalah dengan menggunakan salah satu cabang dari ilmu matematika, yaitu data mining dan logika fuzzy.

Data mining adalah adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam daftar data. Data mining merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai daftar data besar [6]. Dalam data mining terdapat sebuah metode yang digunakan untuk mengklaster data menjadi kelompok-kelompok data, yaitu metode k-means. Beberapa peneliti sebelumnya menggunakan metode k-means untuk mengklaster data karena dalam data mining metode k-means adalah metode pengklasteran yang mudah dipahami dengan algoritma yang cukup mudah [7][8][9]. Selain data mining, terdapat cabang ilmu matematika yang mempunyai metode untuk mengklaster data yaitu logika fuzzy.

Logika fuzzy adalah salah satu cabang ilmu matematika yang mempelajari tentang logika kabur. Dimana logika fuzzy ini memiliki rentang keanggotaan berkisar antara 0 dan 1, berbeda dengan logika klasik yang memiliki rentang keanggotan yang bernilai 0 atau 1[10]. Dalam pengklasteran data, metode fuzzy c-means adalah salah satu metode yang digunakan dalam logika fuzzy. Beberapa peneliti sebelumnya menggunakan metode fuzzy c-means dalam penelitiannya, seperti pengklasifikasian sinyal EEG [11][12], dan analisa klasifikasi status gizi[13]. Dalam jurnal ini akan ditunjukkan perbandingan pengklasteran data iris dengan menggunakan metode k-means dan fuzzy c-means dilihat dari *root mean square error (RMSE)*. *Root mean square error (RMSE)* adalah nilai rata-rata kuadrat dari perbedaan nilai estimasi dengan nilai observasi suatu data. Semakin kecil nilai *RMSE* maka data tersebut semakin valid.

2. Tinjauan Pustaka

2.1 Data Mining

Data mining merupakan proses yang menggunakan teknik statistik, perhitungan, kecerdasan buatan dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar [14]. Dalam data mining terdapat sebuah metode yang digunakan untuk mengklaster data, yaitu k-means. Metode k-means merupakan metode pengklasteran data mining yang sering digunakan peneliti untuk mengklaster data. Dalam metode k-means, data-data yang memiliki karakteristik yang sama diklaster dalam satu kelompok dan data yang memiliki karakteristik yang berbeda dikelompokan dengan kelompok lain yang sesuai dengan karakteristik data tersebut, sehingga data yang berada dalam satu kelompok memiliki tingkat variasi yang kecil [9]. Berikut adalah algoritma dari metode k-means:

(1) Masukkan data yang akan diklaster.
(2) Tentukan jumlah klaster.
(3) Ambil sebarang data sebanyak jumlah klaster secara acak sebagai pusat klaster (sentroid).
(4) Hitung jarak antara data dengan pusat klaster, dengan menggunakan persamaan :

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + \cdots + (X_{ki} - X_{kj})^2} \quad (2.1.1)$$

ISSN: 2527-3159 **E-ISSN: 2527-3167**

Dimana :
$D(i,j)$ = jarak data ke i ke pusat klaster j
X_{ki} = data ke i pada atribut ke k
X_{kj} = titik pusat ke j pada atribut ke k

(5) Hitung kembali pusat klaster dengan keanggotaan klaster yang baru
(6) Jika pusat klaster tidak berubah maka proses klaster telah selesai, jika belum maka ulangi langkah ke (4) sampai pusat klaster tidak berubah lagi.

2.2 Logika Fuzzy

Logika fuzzy pertama kali diperkenalkan oleh Prof. Lotfi A. Zadeh pada tahun 1965. Dalam banyak hal, logika fuzzy digunakan sebagai suatu cara untuk memetakan permasalahan dari *input* menuju ke *output* yang diharapkan. Dalam logika fuzzy terdapat fuzzy *clustering* yang merupakan salah satu metode untuk menentukan klaster optimal dalam suatu ruang vektor yang didasarkan pada bentuk normal *Euclidian* untuk jarak antar vektor[15]. Dalam logika fuzzy terdapat metode yang sering digunakan untuk mengklaster data, yaitu metode fuzzy c-means. Fuzzy c-means adalah suatu metode pengklasteran data yang ditentukan oleh derajat keanggotaan. Berikut adalah algoritma fuzzy c-means:

1. Masukkan data yang akan diklaster, berupa matriks berukuran $n \times m$.
2. Tentukan :
   a. Jumlah klaster = c
   b. Pangkat = w
   c. Maksimum Iterasi = MaxIter;
   d. Error Terkecil yang diharapkan = ε
   e. Fungsi objektif awal = $P_0 = 0$
   f. Iterasi awal = $t = 1$
3. Bangkitkan bilangan acak (μ_{ik}), dengan $i = 1,2,\dots,n; k = 1,2,\dots,c;$ sebagai elemen-elemen matriks partisi awal U.
   Hitung jumlah setiap kolom:
   $$Q_i = \sum_{k=1}^{c} \mu_{ik} \qquad (2.2.1)$$
   dengan $j = 1,2,\dots,n$
   Hitung:
   $$\mu_{ik} = \frac{\mu_{ik}}{Q_i} \qquad (2.2.2)$$
4. Hitung pusat klaster ke- k: V_{kj}
   $$V_{kj} = \frac{\sum_{i=1}^{n}((\mu_{ik})^w * X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^w} \qquad (2.2.3)$$
   dengan $k = 1,2,\dots,c$; dan $j = 1,2,\dots,m$
5. Hitung fungsi objektif pada iterasi ke- t,
   $$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right](\mu_{ik})^w\right) \qquad (2.2.4)$$
6. Hitung perubahan matriks partisi:
   $$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}} \qquad (2.2.5)$$
   dengan $i = 1,2,\dots,n$ dan $k = 1,2,\dots,c$
7. Cek kondisi berhenti:
   a. Jika: $(|P_t - P_{t-1}| < \varepsilon)$ atau $(t > MaxIter)$ maka berhenti,
   b. Jika tidak: $t = t + 1$, ulangi langkah ke-4

Output yang dihasilkan dari *Fuzzy C-Means* (FCM) merupakan deretan pusat klaster dan beberapa derajat keanggotaan untuk tiap-tiap titik data.

2.3 Root Mean Square Error

Root mean ssquare error (RMSE) merupakan parameter yang digunakan untuk mengevaluasi nilai hasil dari pengukuran terhadap nilai sebenarnya atau nilai dianggap benar. Semakin kecil nilai *RMSE*, maka pengklasteran data semakin mendekati benar. Secara umum, persamaan yang digunakan untuk menghitung nilai *RMSE* adalah seperti pada persamaan 2.3.1 sebagai berikut.

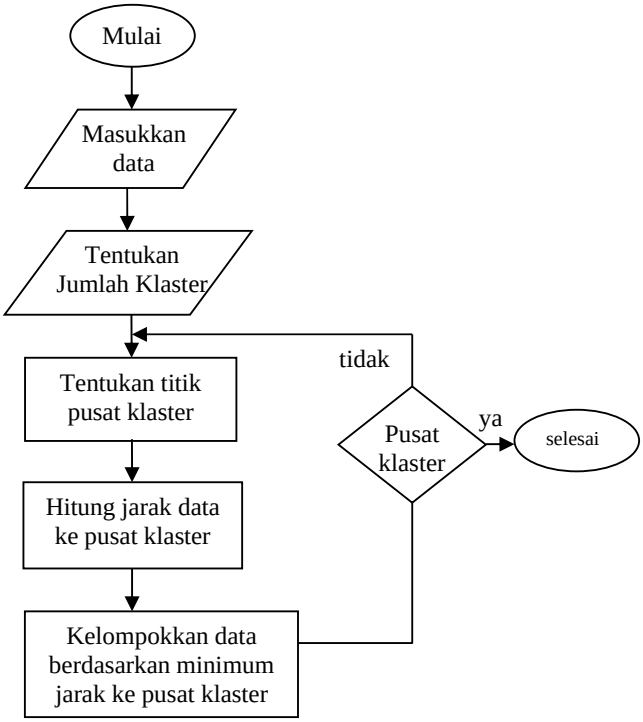
$$RMSE = \sqrt{\frac{(x' - x)^2 + (y' - y)^2}{n}} \qquad (2.3.1)$$

dimana:
(x', y') = nilai perhitungan
(x, y) = nilai exact
n = jumlah data

ISSN: 2527-3159 E-ISSN: 2527-3167

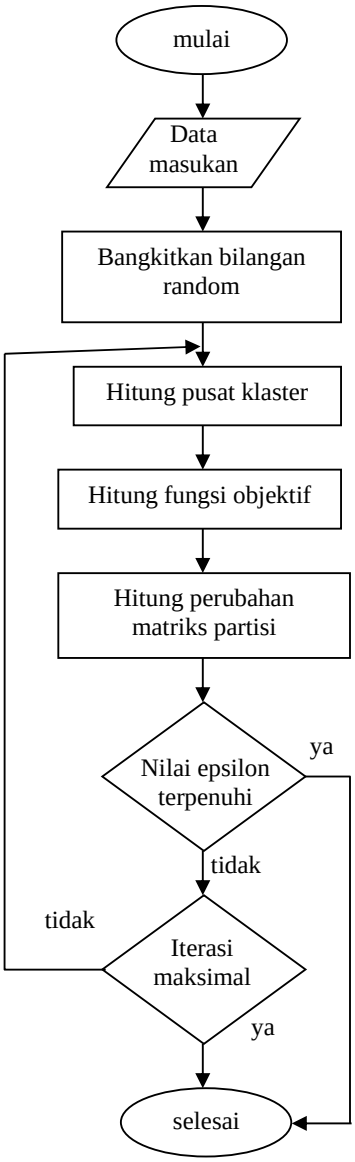
3 Metode Penelitian

Pada jurnal ini, pengklasteran data iris menggunakan dua metode, yaitu metode k-means dan fuzzy c-means. Seperti yang telah dijelaskan pada bab sebelumnya mengenai algoritma dua metode tersebut, terdapat perbedaan pada masing-masing algoritma. Untuk lebih memahami perbedaan kedua algoritma tersebut, dapat dilihat dari *flowchart* algoritma K-Means seperti pada Gambar 3.1.



Gambar 3.1 Algoritma K-Means

Dalam pengklasteran data iris menggunakan metode K-Means, hal yang pertama dilakukan adalah memasukkan data iris terlebih dahulu. Setelah itu, tentukan jumlah klaster yang diharapkan. Lalu tentukan pula titik pusat klaster yang secara acak diambil dari data. Selanjutnya dengan menggunakan persamaan (2.1.1), hitung jarak data ke pusat klaster. Setelah itu, kelompokkan data berdasarkan hasil minimum perhitungan jarak data kepusat klaster. Lalu ulangi lagi langkah awal untuk mengecek apakah titik pusat klaster yang telah dihasilkan sudah tepat dengan mengambil sebarang data dari data baru hasil dari perhitungan jarak data ke pusat klaster. Jika titik pusat klaster berubah maka kita ulangi lagi langkah-langkah sebelumnya sehingga titik pusat klaster tidak berubah.



Gambar 3.2 Algoritma Fuzzy C-Means

Pada gambar 3.2 diatas menunjukkan algoritma pengklasteran data menggunakan metode fuzzy c-means. Sebagai langkah awal yang perlu dilakukan adalah memasukkan data yang akan diklaster dalam bentuk matriks $n \times m$. Lalu tentukan beberapa indikator yang

ISSN: 2527-3159 E-ISSN: 2527-3167

diperlukan pada metode fuzzy c-means. Setelah itu bangkitkan bilangan random dengan menggunakan persamaan 2.2.1. Lalu, hitung pusat klaster dengan menggunakan persamaan 2.2.1. Dari perhitungan pusat klaster, hitung fungsi objektif pada iterasi dengan menggunakan persamaan 2.2.4. setelah itu, hitung perubahan matriks partisi dengan menggunakan persamaan 2.2.5. Lalu, cek kondisi berhenti dengan dilihat dari apakah nilai epsilon yang merupakan salah satu indicator telah terpenuhi atau tidak. Jika sudah terpenuhi maka iterasi selesai, jika iterasi telah maksimal maka kondisi berhenti.

Perbandingan dari metode k-means dan fuzzy c-means tidak benrhenti pada algoritma perhitungannya, akan tetapi perbandingannya terlihat ketika dihitung nilai *RMSE*-nya dengan menggunakan persamaan 2.3.1.

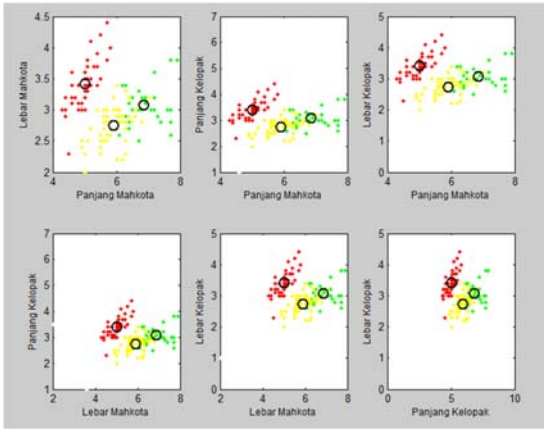
4 Hasil dan Pembahasan

Pada penelitian akan menjelaskan mengenai perbandingan pengklasteran data iris menggunakan metode k-means dan c-means. Akan tetapi, pembahasan ini akan akan direpresentasikan dengan menggunakan *software MATLAB*. Pada MATLAB terdapat fungsi yang dapat digunakan untuk mengklaster data. Pada metode k-means, sebelum mengklaster data menggunakan MATLAB, siapkan data berupa *file* (.dat). setelah itu, tentukan jumlah klaster yangdiharapkan. Lalu, masukkan fungsi metode k-means pada MATLAB, seperti berikut:

```
x=load('datairis.dat');
jumlah_klaster=3;

[center,U,ObjFcn]=fcm(x,jumlah_klaster)
```

Ketika program ini telah disimpan, maka ketika dijalankan akan menghasilkan kelompok-kelompok data. Kelompok-kelompok data tersebut dapat direpresentasikan menggunakan grafik/plot pada MATLAB, sehingga diperoleh sebaran data pada masing-masing klaster berdasarkan titik kedekatannya dengan pusat klaster, hal tersebut terlihat seperti pada Gambar 4.1.



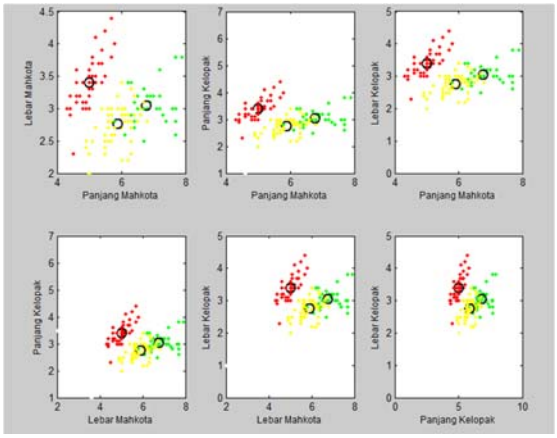
Gambar 4.1 Pengklasteran Iris Menggunakan K-Means

Begitu pula metode fuzzy c-means, metode ini juga menggunakan fungsi pada MATLAB untuk menunjukkan kelompok-kelompok data yang telah diklaster. Adapun fungsi yang digunakan adalah sebagai berikut:

```
x=load('datairis.dat');
jumlah_klaster=3;

[idx,C]=kmeans(x,jumlah_klaster)
```

Ketika fungsi tersebut telah disimpan dan dijalankan akan diperoleh kelompok-kelompok data. Dan dapat juga ditampilkan dengan menggunakan grafik/plot, sehingga diperoleh hasil klasterisasi seperti pada Gambar 4.2.



Gambar 4.2 Pengklasteran Iris Menggunakan Fuzzy C-means

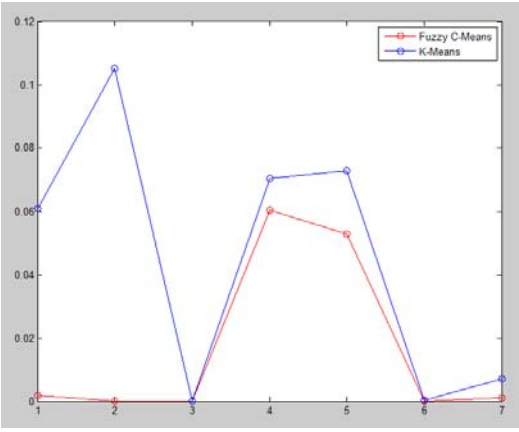
ISSN: 2527-3159 E-ISSN: 2527-3167

Untuk lebih terlihat perbandingan pengklasteran data iris dari kedua metode tersebut, hitung *RMSE* dari data yang sudah diklaster. Perhitungan *RMSE*-pun bisa dilakukan menggunakan MATLAB. Ada beberapa indikator yang harus disiapkan terlebih dahulu, yaitu data *training* dan data *checking*. Data *training* lebih banyak dari data *checking.* Tabel hasil *RMSE* dari dua metode yang berbeda dan data yang sama dapat dilihat pada Tabel 4.1

Tabel 4.1 *RMSE* K-Means dan Fuzzy C-Means

No	Data		Metode	
	Check	Train	FCM	K-Means
1	27	123	0.0530	0.0728
2	35	115	0.0019	0.0608
3	40	160	0.0011	0.0072
4	44	106	0.0604	0.0705
5	60	90	2,2166E-5	0.1051
6	63	87	8,3924E-5	2,6578E-3
7	70	80	2,2122E-14	4,1188E-13

Untuk lebih jelasnya. Perbandingan *RMSE* dari kedua data tersebut dapat direpresentasikan menggunakan grafik/plot, sehingga diperoleh seperti pada Gambar 4.3.



Gambar 4.3 Perbandingan *RMSE* dari K-Means dan Fuzzy C-Means

Dari grafik gambar 4.3, garis biru merepresentasikan hasil perhitungan *RMSE* dari metode fuzzy c-means dan garis hijau merepresentasikan hasil perihtungan *RMSE* dari metode k-means.

5 KESIMPULAN

Dari pembahasan yang telah disampaikan, dapat disimpulkan bahwasanya hasil pengklasteran data iris menggunakan metode k-means dan fuzzzy c-means berbeda. Jika dilihat hasil perhitungan *RMSE* dari kedua metode tersebut, menunjukkan bahwa metode fuzzy c-means memiliki nilai *RMSE* yang lebih kecil dibandingkan dengan nilai *RMSE* metode k-means. Hal ini menunjukkan bahwa pengklasteran menggunakan metode fuzzy c-means lebih mendekati ketepatan (valid) dibandingkan dengan metode k-means.

Penelitian ini masih jauh dari sempurna, masih perlu dilakukan penelitian dengan menggunakan data yang berbeda dan menggunakan lebih banyak data *training* dan *checking* lebih banyak untuk mendapatkan nilai *RMSE.*

6 DAFTAR PUSTAKA

[1] Siregar, Mustaid. Jumlah Spesies Tumbuhan Flora di Indonesia, diambil dari http://www.lipi.go.id/, pada tanggal 28 Juni 2016

[2] Kuniawati, Rizki Taher dkk. Pengelompokan Kualitas Udara Ambien Menurut Kabupaten/Kota di Jawa Tengah Menggunakan Analisis Klaster. Jurnal Gaussian, Vol 4 No 2 Tahun 2015 : 393-402

[3] Kadir, Abdul. Identifikasi Tiga Jenis Bunga iris Menggunakan ANFIS.

[4] Azmi, Meri. Komparasi Metode Jaringan Syaraf Tiruan SOM dan LUQ Untuk Mengidentifikasi Data Bunga Iris. Jurnal TEK NOIF, Vol 2 No 1, April 2014

[5] Riyanto, Hendrik Puasa dkk. Analisa dan Implementasi Fuzzy Inference System Pada Hasil Klasterisasi ALgoritma Fuzzy Subtractive

ISSN: 2527-3159 **E-ISSN: 2527-3167**

Clustering. Universitas TELKOM. 2010

[6] Pane, Dewi Kartika. Implementasi Data Mining Pada Penjualan Produk Elektronik dengan Algoritma Apriori (Studi Kasus : Kreditplus). Jurnal Pelita Informatika Budi Darma, Vol IV No 3, Agustus 2013

[7] Narwati. Pengelompokan Mahasiswa Menggunakan K-Means. Semarang: Fakultas Teknologi Informasi UNISBANK. 2010

[8] Rivani, Edmira. Aplikasi K-Means Cluster Untuk Pengelompokan Provinsi Berdasarkan Produksi Jagung, Padi, Kedelai dan Kacang Hijau. pusat Pengkajian Pengolahan Data dan Informasi, Sekretaris Jenderal DPR RI . Jurusan Statistika Terapan, Universitas Padjajaran, Bandung

[9] Ong, Johan Oscar. Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University. Jurnal Ilmiah Teknik Industri, Vol 12, No 1, Juni 2013 .

[10] Kusumadewi, Sri dan Purnomo, Hari. Aplikasi Logika Fuzzy untuk pendukung keputusan. Edisi 2. Yogyakarta. Graha Ilmu. 2010

[11] Rini, Dian C, Klasifikasi Sinyal EEG Menggunakan Metode Fuzzy C-Means (FCM) Clustering dan Adaptive Neuro Fuzzy Inference System (ANFIS). Undergraduate Thesis, Department of Information Technology, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Indonesia. 2013

[12] Rini, Dian C, Klasifikasi Sinyal EEG Menggunakan Metode Fuzzy C-Means Clustering (FCM) Dan Adaptive Neighborhood Modified Backpropagation (ANMBP). Fakultas Sains dan Teknologi. Universitas Islam Negeri Sunan Ampel Surabaya. 2015.

[13] Sudirman, Nerfita Nikentari dan Martaleli. Analisa Klasifikasi Status Gizi Dengan Metode Fuzzy C-Means Menggunakan Aplikasi Berbasis Android. Jurusan Informatika. Universitas Maritim Raja Ali Haji. Tanjung Pinang

[14]Sutrisno, Afriyudi Wiyanto. Penerapan Data Mining Pada Penjualan Menggunakan Metode Clustering Study Kasus PT. Indoamarco Palembang. Palembang: Universitas Bina Darma.