

Fuzzy C-Mean untuk Mengcluster Pemain Football FIFA (Studi Kasus: Data Kaggle)

Mohammad Afan¹, Retno Wardhani², Nur Nafi'iyah³

^{1,2,3} Program Studi Teknik Informatika, Fakultas Teknik,
Universitas Islam Lamongan
Jalan Veteran Nomor 53A Lamongan
retzno@yahoo.com, mynaff26@gmail.com

Abstrak

Dari kumpulan data yang ada di dunia internet, kami ingin melakukan pengelolaan. Tujuan pengelolaan data yang ada di internet khususnya kaggle adalah untuk mengetahui bagaimana proses clustering serta bagaimana menerapkan algoritma fuzzy c-mean. Fuzzy c-mean digunakan untuk mengcluster dan mendapatkan nilai centroid dari masing-masing cluster. Hasil dari nilai centroid masing-masing cluster digunakan untuk mengelompokkan data baru atau dataset itu sendiri. Kami menggunakan dataset kaggle FIFA Football untuk dicluster menjadi 2 dan 3 kelompok. Algoritma fuzzy c-mean diimplementasikan untuk mencluster data sebanyak 2499 baris dengan 46 variabel. Tool yang digunakan untuk mengolah data adalah matlab.

Kata kunci—FIFA Fottbal, Clustering, Fuzzy C-mean, Matlab

Abstract

From the data set in the internet world, we want to do the management. The purpose of managing existing data on the internet especially Kaggle is to find out how the clustering process is and how to apply the c-mean fuzzy algorithm. Fuzzy c-mean is used to cluster and get the centroid value of each cluster. The results of the centroid value of each cluster are used to group new data or the dataset itself. We use the FIFA Football Kaggle dataset to cluster into 2 and 3 groups. Fuzzy c-mean algorithm is implemented to cluster data with 2499 rows with 46 variables. The tool used to process data is matlab.

Keywords— FIFA Fottbal, Clustering, Fuzzy C-mean, Matlab

1. PENDAHULUAN

Clustering merupakan mengelompokkan data sesuai dengan karakteristik yang sama. *Clustering* dapat digunakan untuk mengkluster data, ataupun mengkluster citra. Salah satu contoh penelitian terkait mengkluster data citra adalah penelitian Nur Nafi'iyah. Melakukan *clustering* citra untuk mengubah citra *grayscale* ke bentuk biner. Di mana proses *clustering* menggunakan algoritma SOM. Setiap intupan piksel dilatih untuk menghasilkan nilai bobot, dan bias. Dari bobot dan bias hasil latih digunakan untuk mengubah citra ke bentuk biner (Nafi'iyah, 2015). Selain digunakan untuk mengkluster data citra, algoritma SOM juga digunakan untuk mengkluster keahlian mahasiswa Teknik Informatika Universitas Islam Lamongan (Nafi'iyah, *Clustering Keahlian Mahasiswa dengan SOM (Studi Khusus: Teknik Informatika Unisla)*, 2015).

Di mana algoritma SOM digunakan untuk mengelompokkan keahlian mahasiswa menjadi 4 jurusan diantaranya keahlian informatika/logika, *software develop/engginer*, manajemen *database*, dan *networking/infrastruktur* dengan akurasi 82%. Variabel yang digunakan untuk mengelompokkan 4 keahlian mahasiswa Teknik Informatika, yaitu: kalkulus 1, kalkulus 2, statistik probabilitas, matematika diskrit, metode numerik, logika matematika, algoritma pemrograman 1, algoritma pemrograman 2, pemrograman visual, pemrograman berorientasi objek, pemrograman internet, rekayasa perangkat lunak, analisa desain berorientasi objek, proyek perangkat lunak, sistem digital, organisasi dan arsitektur komputer, jaringan komputer, struktur data, basis data 1, basis data 2.

Sebagian besar *clustering* digunakan untuk mengelompokkan data yang berupa number atau angka. Akan tetapi tidak hanya data angka, *clustering* juga digunakan untuk mengelompokkan data *text* ataupun citra. Berikut beberapa penelitian terkait pengelompokkan data berupa angka. Penelitian Adi Suryaputra terkait mengkluster data *bandwith* internet. Proses *clustering* menggunakan algoritma Fuzzy C-mean dan Fuzzy C-mean berdasarkan pemilihan fitur yang berkorelasi. Untuk memilih fitur yang berkorelasi menggunakan metode *correlation based feature selection*. Fitur-fitur yang dipilih diantaranya: *flow metrics*, *packet inter-arrival time*, *size of TOP/IP control fields*, *total packets*, *payload size*, *effective bandwidth based upon entropy*, *ranked list of top-ten*, dan lainnya. Hasil akurasi dalam mengkluster data terdapat perbedaan, jika menggunakan Fuzzy C-mean saja akurasinya 83,73%. Sedangkan jika menggunakan pemilihan korelasi fitur hasil *clustering* akurasinya 88,10% (Adi Suryaputra P, Febriliyan Samopa, Bekti Cahyo Hindayanto, 2014).

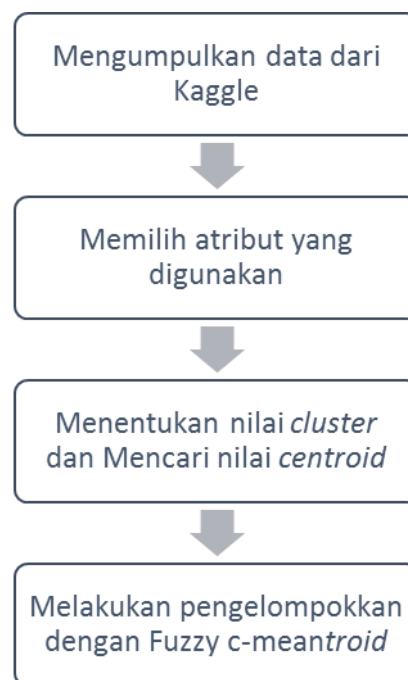
Penelitian lainnya yang terkait pengelompokkan data numerik adalah Pengelompokkan Pasar. Di mana melakukan segmentasi pasar agar dapat mengetahui perilaku konsumen dan bagaimana daya minat membeli konsumen. Tujuan penelitian melakukan segmentasi pasar untuk dapat memudahkan membuat strategi pemasaran. Atribut yang digunakan untuk mensegmentasi pasar adalah *recency*, *frequency* dan *monetary*. Proses *clustering* menggunakan algoritma FCM dan *Fuzzy Subtractive Clustering* dengan *tool* matlab (Titus Kristanto, Rachman Arief, 2013). Penelitian lain yang terkait *clustering* data numerik nilai siswa dan mahasiswa adalah (Tb. Ai Munandar, Wahyu Oktri Widyarto, Harsiti, 2013), (Nelson Butarbutar, Agus Perdana Windarto, Dedi Hartama, Solikhun, 2016), (Cary Lineker Simbolon, Nilamsari Kusumastuti, Beni Irawan, 2013). Ketiga penelitian tersebut terkait mengelompokkan nilai baik siswa ataupun mahasiswa. Penelitian Ali Munandar mengolah data nilai mahasiswa untuk mengelompokkan konsentrasi jurusan mahasiswa. Di mana terdapat 3 kelompok, yaitu: konsentrasi multimedia, konsentrasi web, dan konsentrasi pemrograman. Variabel yang digunakan untuk mengelompokkan adalah pemrograman internet, algoritma pemrograman, pemrograman terstruktur, komunikasi data, struktur data, matematika diskrit, IP semester 1, IP semester 2, IP semester 3, dan IP semester 4. Penelitian lainnya yang mengelompokkan nilai mahasiswa adalah Cary Lineker, yaitu mengelompokkan lulusan mahasiswa Matematika Universitas Tangjungpura berdasarkan IPK dan massa studi. Pengelompokkan sebanyak 4 kelompok, dengan kelompok 1 IPK 3.15 massa studi 5.09 tahun, kelompok 2 IPK 2.88 massa studi 7.32 tahun, kelompok 3 IPK 3.48 massa studi 4.37 tahun, dan kelompok ke-4 IPK 2.89 massa studi 5.91 tahun. Hasil penelitian Cary menunjukkan kelompok ke-4 merupakan kelompok yang paling banyak dengan jumlah lulusan 33 mahasiswa. Penelitian lain terkait pengelompokkan, yaitu Nelson Butarbutar dalam mengkluster nilai siswa untuk melihat prestasi akademik. *Clustering* siswa berdasarkan nilai rata-rata raport yang

dikelompokkan menjadi 3, yaitu kelompok ke-1 nilai rata-rata lebih dari sama dengan 7.24, kelompok ke-2 nilai rata-rata antara 6-7.24, dan kelompok ke-3 nilai rata-rata kurang dari sama dengan 4. Dari ketiga penelitian di atas (Tb. Ai Munandar, Wahyu Oktri Widyarto, Harsiti, 2013), (Nelson Butarbutar, Agus Perdana Windarto, Dedi Hartama, Solikhun, 2016), (Cary Lineker Simbolon, Nilamsari Kusumastuti, Beni Irawan, 2013) menggunakan algoritma *Fuzzy C-means*.

Dari kondisi penelitian sebelumnya yang melakukan *clustering* dengan menggunakan metode *fuzzy C-mean*. Maka kami ingin membuat *clustering* pemain football FIFA data dari *Kaggle* dengan algoritma *Fuzzy C-mean*.

2. METODE PENELITIAN

Alur penelitian ini seperti Gambar 1.



Gambar 1. Langkah Penelitian

Data diambil dari dataset Kaggle FIFA football. Data terdapat puluhan ribu variabel. Akan tetapi agar dapat digunakan untuk mengelompokkan data pemain football FIFA kami hanya menggunakan variabel yang berupa number/angka seperti Tabel 1. Variabel yang digunakan untuk pengelompokkan hanya 46 variabel, diambil data yang number. Data yang digunakan hanya 2499 baris dari total *dataset* 18205 baris.

Tabel 1. Variabel yang Digunakan Mengelompokkan

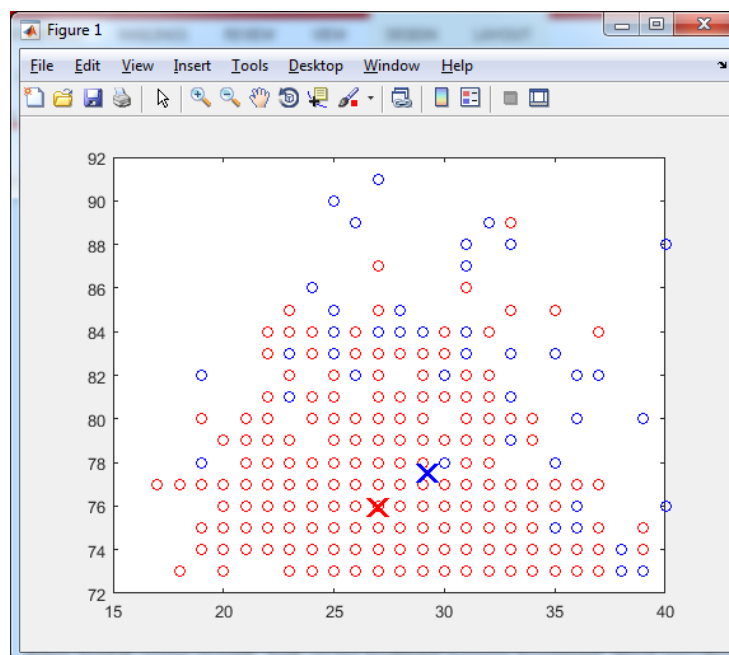
| Nomor | Nama Variabel |
|-------|---------------|
| 1 | Age |
| 2 | Overall |
| 3 | Potential |

| Nomor | Nama Variabel |
|-------|---------------------------------|
| 4 | <i>Value</i> |
| 5 | <i>Wage</i> |
| 6 | <i>Special</i> |
| 7 | <i>International Reputation</i> |
| 8 | <i>Wea Foot</i> |
| 9 | <i>Sill oves</i> |
| 10 | <i>Jersey Nuber</i> |
| 11 | <i>Weight</i> |
| 12 | <i>Crossing</i> |
| 13 | <i>Finishing</i> |
| 14 | <i>HeadingAccuracy</i> |
| 15 | <i>ShortPassing</i> |
| 16 | <i>Volleys</i> |
| 17 | <i>Dribbling</i> |
| 18 | <i>Curve</i> |
| 19 | <i>FAccuracy</i> |
| 20 | <i>LongPassing</i> |
| 21 | <i>BallControl</i> |
| 22 | <i>Acceleration</i> |
| 23 | <i>SprintSpeed</i> |
| 24 | <i>Agility</i> |
| 25 | <i>Reactions</i> |
| 26 | <i>Balance</i> |
| 27 | <i>ShotPower</i> |
| 28 | <i>Juping</i> |
| 29 | <i>Staina</i> |
| 30 | <i>Strength</i> |
| 31 | <i>LongShots</i> |
| 32 | <i>Aggression</i> |
| 33 | <i>Interceptions</i> |
| 34 | <i>Positioning</i> |
| 35 | <i>Vision</i> |
| 36 | <i>Penalties</i> |
| 37 | <i>Coposure</i> |
| 38 | <i>aring</i> |
| 39 | <i>StandingTacle</i> |
| 40 | <i>SlidingTacle</i> |
| 41 | <i>GDivng</i> |
| 42 | <i>GHandling</i> |
| 43 | <i>Gicing</i> |
| 44 | <i>GPositioning</i> |

| Nomor | Nama Variabel |
|-------|-----------------------|
| 45 | <i>GReflexes</i> |
| 46 | <i>Release Clause</i> |

3. HASIL DAN PEMBAHASAN

Kami melakukan ujicoba untuk mengelompokkan data, ujicoba yang pertama menggunakan 3 kelompok, hasil pengelompokkan seperti Gambar 2. Dalam melakukan *clustering* terjadi 43 iterasi, hasil nilai *centroid* dari ke-3 *cluster* dalam Tabel 2.

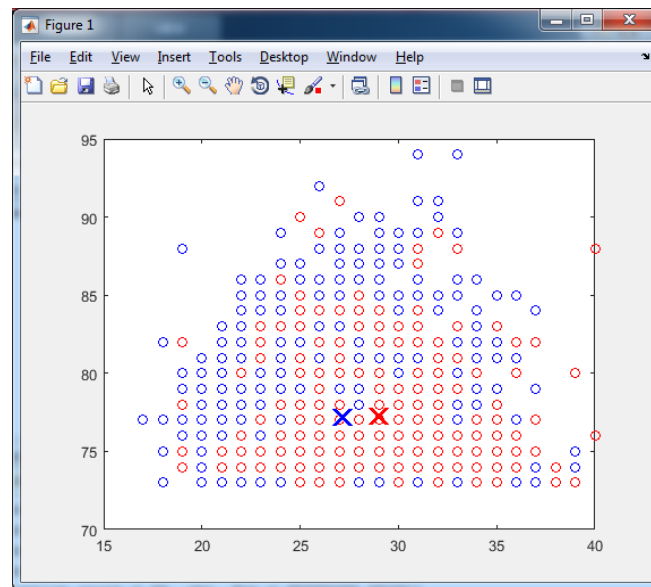


Gambar 2. Hasil Pengelompokkan dengan 3 Cluster

Tabel 2. Nilai Centroid dari 3 Cluster

| | | | |
|----|----------|----------|----------|
| 1 | 29.22338 | 26.9811 | 27.26395 |
| 2 | 77.53085 | 75.94702 | 77.84672 |
| 3 | 79.16544 | 78.32336 | 79.73092 |
| 4 | 19.59631 | 11.59967 | 14.34663 |
| 5 | 30.32982 | 30.60675 | 44.60566 |
| 6 | 1227.774 | 1795.971 | 2008.951 |
| 7 | 1.607041 | 1.423752 | 1.754507 |
| 8 | 2.643474 | 3.101886 | 3.315345 |
| 9 | 1.053505 | 2.73613 | 3.282717 |
| 10 | 13.95017 | 15.43923 | 15.26503 |
| 11 | 185.79 | 171.5422 | 163.9441 |
| 12 | 16.1266 | 57.23928 | 71.10886 |

| | | | |
|----|----------|----------|----------|
| 13 | 13.87633 | 54.36986 | 63.87622 |
| 14 | 17.7548 | 67.83238 | 62.93703 |
| 15 | 32.80527 | 70.12443 | 76.78495 |
| 16 | 14.66123 | 51.31113 | 62.73651 |
| 17 | 16.84487 | 65.8087 | 75.63402 |
| 18 | 16.69641 | 55.30667 | 70.30246 |
| 19 | 15.69514 | 48.75494 | 63.87078 |
| 20 | 31.74054 | 62.33278 | 71.65342 |
| 21 | 24.14213 | 70.58047 | 77.50877 |
| 22 | 46.16654 | 67.41213 | 73.74782 |
| 23 | 46.56541 | 68.83792 | 72.97019 |
| 24 | 48.66805 | 66.06196 | 74.64847 |
| 25 | 72.92337 | 72.37305 | 75.79947 |
| 26 | 44.66877 | 63.58191 | 72.77258 |
| 27 | 25.62106 | 66.16312 | 74.87853 |
| 28 | 66.6016 | 69.6374 | 69.79471 |
| 29 | 36.14463 | 70.34594 | 77.69284 |
| 30 | 67.93701 | 72.9303 | 68.89548 |
| 31 | 15.13628 | 56.01538 | 69.74768 |
| 32 | 32.46932 | 67.5856 | 70.75619 |
| 33 | 22.57363 | 56.08019 | 63.85807 |
| 34 | 13.86278 | 58.80638 | 70.65099 |
| 35 | 47.09278 | 60.04875 | 72.45128 |
| 36 | 22.70397 | 55.98013 | 63.81204 |
| 37 | 57.35623 | 70.82917 | 74.83544 |
| 38 | 19.46013 | 56.51421 | 62.51521 |
| 39 | 17.13813 | 56.5284 | 63.71419 |
| 40 | 16.87088 | 53.41207 | 60.33252 |
| 41 | 75.01121 | 10.79921 | 10.66317 |
| 42 | 72.0218 | 10.83592 | 10.73952 |
| 43 | 68.32055 | 10.75289 | 10.75599 |
| 44 | 73.45326 | 10.789 | 10.72763 |
| 45 | 76.01938 | 10.81893 | 10.61373 |
| 46 | 22.51183 | 17.46339 | 24.77499 |



Gambar 3. Hasil *Clustering 2 Cluster*

Ujicoba yang kedua menggunakan 2 kelompok, hasil pengelompokan seperti Gambar 3. Dalam melakukan *clustering* terjadi 33 iterasi, hasil nilai centroid dari ke-2 *cluster* dalam Tabel 3.

Tabel 3. Nilai Centroid Hasil *2 Cluster*

| | <i>Cluster 1</i> | <i>Cluster 2</i> |
|----|------------------|------------------|
| 1 | 27.15065 | 28.98279 |
| 2 | 77.18029 | 77.276 |
| 3 | 79.24048 | 79.00501 |
| 4 | 13.26384 | 26.45208 |
| 5 | 40.10303 | 31.67624 |
| 6 | 1935.871 | 1305.175 |
| 7 | 1.641018 | 1.593409 |
| 8 | 3.254964 | 2.682281 |
| 9 | 3.128763 | 1.261026 |
| 10 | 15.56202 | 13.6995 |
| 11 | 165.9521 | 185.2252 |
| 12 | 66.79992 | 20.65524 |
| 13 | 61.52523 | 17.91943 |
| 14 | 64.27171 | 28.55112 |
| 15 | 74.55304 | 39.03733 |
| 16 | 59.49521 | 18.34969 |
| 17 | 72.90824 | 23.2227 |
| 18 | 65.76498 | 20.75679 |
| 19 | 59.0513 | 19.30962 |

| | <i>Cluster 1</i> | <i>Cluster 2</i> |
|----|------------------|------------------|
| 20 | 68.29436 | 36.9498 |
| 21 | 75.47289 | 31.41837 |
| 22 | 72.21341 | 47.9894 |
| 23 | 72.10888 | 48.86454 |
| 24 | 72.43539 | 49.19596 |
| 25 | 74.61986 | 72.69364 |
| 26 | 70.17036 | 45.8475 |
| 27 | 72.20178 | 31.38628 |
| 28 | 69.61635 | 67.20539 |
| 29 | 75.05112 | 41.67885 |
| 30 | 69.72244 | 70.52476 |
| 31 | 65.59986 | 19.31421 |
| 32 | 68.85929 | 40.60445 |
| 33 | 59.49776 | 31.59733 |
| 34 | 67.46687 | 18.52909 |
| 35 | 68.67186 | 46.97305 |
| 36 | 61.5678 | 26.82992 |
| 37 | 73.50109 | 59.6862 |
| 38 | 58.93592 | 29.07685 |
| 39 | 59.44294 | 27.41744 |
| 40 | 56.08409 | 26.65801 |
| 41 | 10.66721 | 63.00146 |
| 42 | 10.73577 | 60.49527 |
| 43 | 10.74126 | 57.56953 |
| 44 | 10.68731 | 61.75507 |
| 45 | 10.63684 | 63.79652 |
| 46 | 22.33122 | 25.3111 |

Hasil pengelompokkan dengan menggunakan 2 *cluster* ataupun 3 *cluster* dalam Tabel 4.

Tabel 4. Hasil Pengelompokkan

| Data | <i>2-Cluster</i> | <i>3-Cluster</i> |
|------|------------------|------------------|
| 1 | <i>Cluster 1</i> | <i>Cluster 2</i> |
| 2 | <i>Cluster 1</i> | <i>Cluster 2</i> |
| 3 | <i>Cluster 1</i> | <i>Cluster 2</i> |
| 4 | <i>Cluster 1</i> | <i>Cluster 2</i> |
| 5 | <i>Cluster 1</i> | <i>Cluster 2</i> |
| 6 | <i>Cluster 1</i> | <i>Cluster 2</i> |
| 7 | <i>Cluster 1</i> | <i>Cluster 2</i> |
| 8 | <i>Cluster 1</i> | <i>Cluster 2</i> |
| 9 | <i>Cluster 1</i> | <i>Cluster 2</i> |

| Data | 2-Cluster | 3-Cluster |
|------|-----------|-----------|
| 10 | Cluster 1 | Cluster 2 |
| 11 | Cluster 1 | Cluster 2 |
| 12 | Cluster 1 | Cluster 2 |
| 13 | Cluster 1 | Cluster 2 |
| 14 | Cluster 1 | Cluster 2 |
| 15 | Cluster 1 | Cluster 2 |
| 16 | Cluster 1 | Cluster 2 |
| 17 | Cluster 1 | Cluster 2 |
| 18 | Cluster 1 | Cluster 2 |
| 19 | Cluster 1 | Cluster 2 |
| 20 | Cluster 1 | Cluster 2 |
| 21 | Cluster 1 | Cluster 2 |
| 22 | Cluster 1 | Cluster 2 |
| 23 | Cluster 1 | Cluster 2 |
| 24 | Cluster 1 | Cluster 2 |
| 25 | Cluster 1 | Cluster 2 |
| 26 | Cluster 1 | Cluster 2 |
| 27 | Cluster 1 | Cluster 2 |
| 28 | Cluster 1 | Cluster 2 |
| 29 | Cluster 1 | Cluster 2 |
| 30 | Cluster 1 | Cluster 2 |
| 31 | Cluster 1 | Cluster 2 |
| 32 | Cluster 1 | Cluster 2 |
| 33 | Cluster 1 | Cluster 2 |
| 34 | Cluster 1 | Cluster 2 |
| 35 | Cluster 1 | Cluster 2 |
| 36 | Cluster 1 | Cluster 2 |
| 37 | Cluster 1 | Cluster 2 |
| 38 | Cluster 1 | Cluster 2 |
| 39 | Cluster 1 | Cluster 2 |
| 40 | Cluster 2 | Cluster 1 |
| 41 | Cluster 1 | Cluster 2 |
| 42 | Cluster 1 | Cluster 2 |
| 43 | Cluster 1 | Cluster 2 |
| 44 | Cluster 1 | Cluster 2 |
| 45 | Cluster 1 | Cluster 2 |
| 46 | Cluster 1 | Cluster 2 |
| 47 | Cluster 1 | Cluster 2 |
| 48 | Cluster 1 | Cluster 2 |
| 49 | Cluster 1 | Cluster 2 |
| 50 | Cluster 1 | Cluster 2 |

4. KESIMPULAN

Dari kegiatan di atas mulai dari memilih variabel dan mengelompokkan data. Maka kesimpulan yang dapat disampaikan adalah proses *clustering* baik menggunakan 2 kelompok ataupun 3 kelompok cukup baik. Dilihat dari hasil *cluster* memang terdapat perbedaan, karena setiap kita melakukan uji coba untuk mencari nilai centroid akan dihasilkan nilai random. Kegiatan *clustering* ini menggunakan *tool* matlab dan fungsi *fuzzy c-mean* matlab.

5. SARAN

Dapat mengimplementasikan *fuzzy c-mean* untuk mengcluster data yang lain dari Kaggle dan kesemua dataset harus diolah serta digunakan secara tepat.

DAFTAR PUSTAKA

- Adi Suryaputra P, Febriliyan Samopa, Bekti Cahyo Hindayanto. (2014). *Klasterisasi dan Analisis Trafik Internet Menggunakan Fuzzy C Mean Dengan Ekstraksi Fitur Data*. *Jurnal Informatika*, 33-39.
- Cary Lineker Simbolon, Nilamsari Kusumastuti, Beni Irawan. (2013). *Clustering Lulusan Mahasiswa Matematika FMIPA Untan Pontianak Menggunakan Algoritma Fuzzy C-means*. *Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster)*, 21-26.
- Nafi'iyah, N. (2015). *Algoritma Kohonen dalam Mengubah Citra Graylevel Menjadi Citra Biner*. *Jurnal Ilmiah Teknologi dan Informasia ASIA (JITIKA)*, 49-55.
- Nafi'iyah, N. (2015). *Clustering Keahlian Mahasiswa dengan SOM (Studi Khusus: Teknik Informatika Unisla)*. *SNATIKA* (pp. 105-110). Malang: STIKI Indonesia Malang.
- Nelson Butarbutar, Agus Perdana Windarto, Dedi Hartama, Solikhun. (2016). *Komparasi Kinerja Algoritma Fuzzy C-Means dan K-Means Dalam Pengelompokan Data Siswa Berdasarkan Prestasi Nilai Akademik Siswa*. *JURASIK (Jurnal Riset Sistem Informasi & Teknik Informatika)*, 46-55.
- Tb. Ai Munandar, Wahyu Oktri Widyarto, Harsiti. (2013). *Clustering Data Nilai Mahasiswa untuk Pengelompokan Konsentrasi Jurusan Menggunakan Fuzzy Cluster Means*. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, (pp. 30-33). Yogyakarta.
- Titus Kristanto, Rachman Arief. (2013). *Analisa Data Mining Metode Fuzzy untuk Customer Relationship Management pada Perusahaan Tour & Travel*. *Seminar Nasional Sistem Informasi Indonesia* (pp. 528-534). Surabaya: SESINDO.