PERBANDINGAN *FUZZY C-MEANS* DAN *K-MEANS* UNTUK MENGELOMPOKKAN TINGKAT BUTA HURUF BERDASARKAN PROVINSI DI INDONESIA

Murtadlo Anugerah Pamungkas[1], Hardian Oktavianto[2], Reni Umilasari[3]

Program Studi Teknik Informatika, Universitas Muhammadiyah Jember
Jl. Karimata No 49, Jember, Jawa Timur
Telp. (+62)82233934478
Email : pamungkasmurtadlo@gmail.com[1], hardian@unmuhjember.ac.id[2], reni.umilasari@gmail.com[3]

Abstrak

Buta huruf adalah ketidakmampuan seseorang dalam membaca atau menulis kalimat sederhana. Buta huruf merupakan salah satu permasalahan dalam menempuh pendidikan dan menghambat proses pencapaian tujuan peningkatan kualitas pendidikan. Permasalahan buta huruf adalah salah satu pekerjaan bagi pemerintah Indonesia dan tantangan untuk dunia pendidikan. Penelitian ini melakukan perbandingan antara algoritma *Fuzzy C-Means* dan algoritma *K-Means* untuk mengetahui algoritma mana yang lebih optimum dalam melakukan metode *clustering*, dengan mencari *cluster* optimum dari masing-masing algoritma dan melakukan uji validitas yang digunakan untuk melakukan perbandingan antara algoritma *Fuzy C-Means* dan *K-Means*. Hasil penelitian ini menunjukkan *cluster* optimum untuk algoritma *Fuzzy C-Means* berada pada 4 *cluster* dan *cluster* optimum untuk algoritma *K-Means* berada pada 6 *cluster*. Perbandingan dilakukan berdasarkan *cluster* optimum dengan menggunakan metode *Partition Coefficient Index* untuk algoritma *Fuzzy C-Means* dengan nilai validitas 0,7684 dan metode *Silhouette Index* untuk algoritma *K-Means* dengan nilai validitas 0,4966. Dapat disimpulkan bahwa algoritma *Fuzzy C-Means* lebih optimum dalam melakukan *clustering* dibandingkan dengan algoritma *K-Means* karena nilai validitas algoritma *Fuzzy C-Means* lebih besar atau mendekati nilai 1.

Kata Kunci: Buta Huruf, *Clustering*, Perbandingan, *Fuzzy C-Means*, *K-Means*, *Dunn Index*, *Partition Coefficient Index*, *Silhouette Index*.

1. Pendahuluan

Bekal dasar dalam menempuh pendidikan adalah sanggup dalam menulis, membaca, berkomunikasi dan berhitung. Buta huruf adalah salah satu permasalahan dalam menempuh pendidikan dan menghambat proses pencapaian tujuan peningkatan kualitas pendidikan, sedangkan yang dimaksud dengan buta huruf yaitu ketidakmampuan seseorang dalam menulis atau membaca kalimat sederhana dalam bahasa Indonesia atau berbagai bahasa lainnya. Dampak dari tidak meratanya pendidikan salah satunya merupakan buta huruf yang menyebabkan rendahnya kualitas sumber daya manusia (SDM).

Permasalahan buta huruf adalah salah satu pekerjaan rumah bagi pemerintahan Indonesia dan tantangan untuk dunia pendidikan. Metode yang dapat menyelesaikan permasalahan tersebut secara kompleks adalah melalui salah satu cabang ilmu komputer seperti *data mining* dengan memakai metode *clustering*. Istilah *data mining* dipakai untuk menjabarkan penemuan pengetahuan di dalam daftar data untuk mengambil dan mengenali informasi yang berguna dari berbagai daftar yang besar. Beberapa algoritma yang terdapat dalam metode *clustering* diantaranya yaitu algoritma *Fuzzy C-Means* dan algoritma *K-Means*. Kedua algoritma tersebut akan mengelompokkan data ke dalam bentuk satu

atau lebih *cluster*. Data yang mempunyai keunikan yang sama akan dikelompokkan ke dalam ke dalam *cluster* yang sejenis dan data yang memiliki keunikan yang berbeda akan dikelompokkan ke dalam *cluster* yang lainnya.

Sebelumnya penelitian telah dilaksanakan oleh (Rahayu, 2019) dengan studi kasus "Analisa *K-Medoids* Dalam Pengelompokan Penduduk Buta Huruf Menurut Provinsi". Penelitian tersebut bertujuan untuk mengimplementasikan algoritma *K-Medoids* sebanyak 3 *cluster*. Pada penelitian yang lain dilakukan oleh (Agustina dan Prihandoko, 2018) dengan studi kasus "Perbandingan Algoritma *K-Means* Dengan Algoritma *Fuzzy C-Means* Untuk *Clustering* Tingkat Kedisiplinan Kinerja Karyawan". Pada penelitian tersebut, tujuan peneliti yaitu melakukan perbandingan algoritma *K-Means* dengan algoritma *Fuzzy C-Means* untuk mengetahui kelebihan dari masing-masing algoritma saat melakukan *clustering*, kemudian memperoleh hasil yang optimum. Hasil dari penelitian tersebut algoritma *Fuzzy C-Means* lebih optimal dalam melakukan *clustering* dengan nilai validitas sebesar 0,758 disebabkan nilai validasi yang mendekati 1, sedangkan algoritma *K-Means* memiliki nilai validasi sebesar 0,528.

Penelitian ini bertujuan untuk mencari *cluster* optimum dari algoritma *Fuzzy C-Means* dan algoritma *K-Means* menggunakan metode *Dunn Index* dengan skenario 3 sampai 10 *cluster*, kemudian dilakukan perbandingan antara kedua algoritma tersebut menggunakan uji validitas berdasarkan *cluster* optimum dari masing-masing algoritma. Melakukan perbandingan algoritma *Fuzzy C-Means* dan *K-Means* untuk melihat algoritma yang lebih baik dalam melakukan *clustering*. Algoritma *K-Means* melakukan uji validitas dengan metode *Silhouette Index* dan algoritma *Fuzzy C-Means* melakukan uji validitas memakai metode *Partitioan Coefficient Index*.

2. Studi Pustaka

2.1 *Data Mining*

Data mining merupakan proses penjabaran data dari sudut pandang yang berbeda dan menjadikannya sebagai informasi yang penting dan bisa digunakan dalam mengurangi biaya pengeluaran, menaikkan keuntungan atau bahkan keduanya. Secara teknis, *data mining* itu sendiri juga disebut sebagai jalan untuk menemukan kedekatan dari banyaknya sekumpulan data yang ada pada sebuah relasional *database* yang besar (Berry dan Linoff, 2004).

2.2 *Clustering*

Clustering adalah suatu proses pengelompokan data yang akan menghasilkan informasi dan menjabarkan relasi objek satu dengan objek lainnya menggunakan prinsip untuk meminimumkan kemiripan antar *cluster* dan memaksimalkan kemiripan antar anggota satu *cluster* yang bertujuan untuk mencari *cluster* yang berkualitas dalam waktu yang layak (Hartigan, 1975).

2.3 *Fuzzy C-Means*

Fuzzy C-Means pertama kali diperkenalkan oleh Jim Bezdek di tahun 1981. Pada metode *clustering* terdapat salah satu teknik yaitu *Fuzzy C-Means* yang berperan dengan mengelompokkan setiap data dalam satu *cluster* yang ditentukan oleh nilai derajat keanggotaan (Gelley, 2000). Menurut (Kusumadewi dan Purnomo, 2004). Langkah-langkah algoritma *Fuzzy C-Means* yaitu sebagai berikut:

1. Masukkan data yang akan di *cluster*.
2. Menentukan:
   a. Jumlah *cluster* (c);
   b. Pangkat (w);
   c. Maksimum iterasi (*MaxIter*);
   d. *Error* terkecil (ξ);
   e. Fungsi objektif awal ($P_0 = 0$);
   f. Iterasi awal (t = 1).
3. Membangkitkan bilangan *random* sebagai derajat keanggotaan awal (μ)

4. Menghitung pusat *cluster* ke-k: V_{kj}, dengan k = 1,2,...,c; dan j = 1,2,...,m.

$$V_{kj} = \frac{\sum_{i-1}^{n}((\mu_{ik})^{w} \times X_{ij})}{\sum_{i-1}^{n}(\mu_{ik})^{w}} \quad (1.1)$$

5. Menghitung fungsi objektif pada iterasi ke-t.

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sqrt{\sum_{j=1}^{m}(X_{ij} - V_{kj})^2}\right](\mu_{ik})^{w}\right) \quad (1.2)$$

6. Menghitung perubahan matriks partisi.

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(X_{ij} - V_{kj})^2\right]^{\frac{-1}{w-1}}} \quad (1.3)$$

7. Memeriksa kondisi berhenti.
   a. Jika (|Pt – Pt-1|< ξ) atau (t > MaxIter) maka berhenti.
   b. Jika tidak t = t + 1, mengulang ke langkah empat.

2.4 *K-Means*

Awal diperkenalkan *K-Means* pada tahun 1957 oleh Stuart P. Llyod, tetapi metode tersebut baru dikemukakan pada tahun 1982. Metode *K-Means clustering* akan mengelompokkan objek atau data ke dalam sejumlah *cluster* (Everitt, 2011). Berikut proses pada algoritma K-Means:

1. Memasukkan data yang akan di *cluster*.
2. Menentukan jumlah *cluster*.
3. Menetapkan *centroid* awal yang diambil secara *random*.
4. Menghitung jarak setiap data ke *centroid*.

$$d(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (2.1)$$

5. Mengelompokkan setiap data berdasarkan jarak minimum dengan *centroid*.

$$Min\sum_{k=i}^{k} d(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \quad (2.2)$$

6. Mencari nilai titik pusat (*centroid*) terbaru.

$$C_{ij} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad (2.3)$$

7. Memerika kondisi berhenti.
   a. *Centroid* baru tidak mengalami perubahan dan memiliki nilai yang sama dengan *centroid* awal.
   b. Anggota *cluster* tidak mengalami perpindahan.

2.5 *Dunn Index*

Dunn Index merupakan metrik untuk mengevaluasi algoritma *clustering*. Tujuan dari metode ini adalah untuk menghitung nilai *compactness* dan nilai *separation*. *Cluster* optimum ditunjukkan dengan semakin besar nilai *Dunn Index* (Dey, 2019).

$$(C) = \min_{1\le i\le c}\left\{\min_{1\le j\le c, j\ne i}\left\{\frac{dist\,(x_i, x_j)}{\max_{1\le k\le c}\{diam\,(x_k)\}}\right\}\right\} \quad (3.1)$$

2.6 *Partition Coefficient Index*

Metode *Partition Coefficient Index* hanya mengoreksi nilai derajat keanggotaan yang biasanya mengandung informasi geometrik. Nilai dalam rentang [0,1], Kualitas *cluster* yang semakin baik akan ditunjukkan dengan nilai yang semakin tinggi atau mendekati nilai 1 (Prasetyo, 2014).

$$(C) = \frac{1}{N}\sum_{i=1}^{C}\sum_{j}^{N}\left(\mu_{ij}\right)^2 \quad (4.1)$$

2.7 *Silhouette Index*

Silhouette Index digunakan untuk memperkirakan seberapa baik suatu penelitian menjadi satu *cluster* atau mengukur nilai rata-rata jarak antar *cluster* untuk melihat kualitas *cluster*. Sebuah *cluster* dinyatakan semakin baik jika memiliki nilai yang semakin mendekati 1 (Izzadin, 2019).

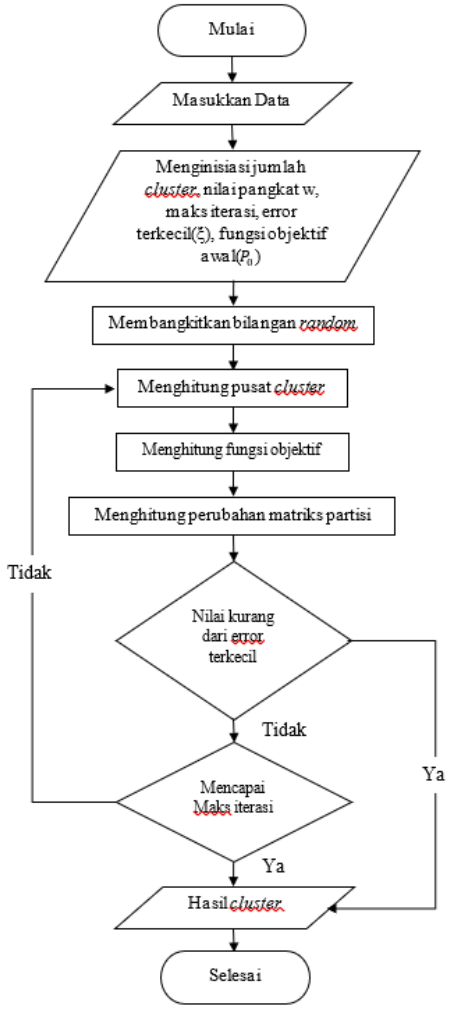
$$SI = \frac{1}{N}\sum_{i=1}^{N} S(x_i) \quad (5.1)$$

3. Metodologi Penelitian

3.1 Pengumpulan Data

Penelitian ini menggunakan data yang didapat dari situs resmi Badan Pusat Statistik Provinsi Jawa Timur. Dataset yang digunakan pada penelitian ini ialah tingkat penduduk buta huruf berdasarkan provinsi di Indonesia. Data tersebut sebanyak 165 data dan terdiri dari 33 provinsi.

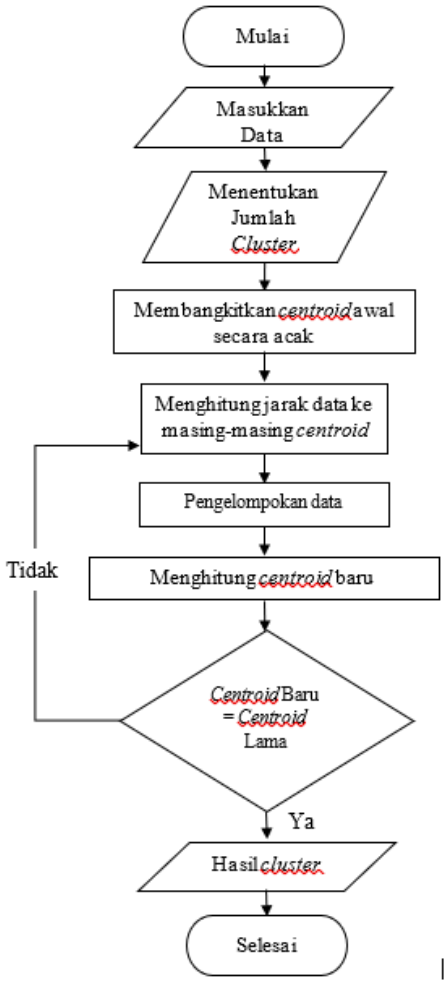
3.2 *Fuzzy C-Means*



Gambar 1 Algoritma *Fuzzy C-Means*

Dalam melakukan *clustering* data memakai algoritma *Fuzzy C-Means*, langkah pertama yang dilakukan yaitu memasukkan data yang akan di *cluster*. Lalu menginisiasi beberapa parameter yang dibutuhkan pada algoritma *Fuzzy C-Means*. Kemudian membangkitkan bilangan *random* sebagai derajat keanggotaan awal. Setelah itu menghitung pusat *cluster* dengan persamaan (1). Di lanjutkan menghitung fungsi objektif menggunakan persamaan (2). Lalu menghitung perubahan matriks partisi menggunakan persamaan (3). Proses selanjutnya melakukan pemeriksaan kondisi berhenti, jika sudah mencapai error terkecil dan maksimal iterasi maka proses iterasi akan berhenti.

3.3 *K-Means*



Gambar 2 Algoritma *K-Means*

Clustering data menggunakan algorimta *K-Means*, langkah pertama yang dilakukan yaitu memasukkan data yang akan di *cluster*. Lalu membangkitkan *centroid* awal secara *random*. Kemudian menghitung jarak data ke setiap *centroid* dengan persamaan (1). Setelah itu, melakukan pengelompokan data dengan persamaan (2). Di lanjutkan dengan menghitung *centroid*

baru menggunakan persamaan (3). Proses selanjutnya melakukan pemeriksaan kondisi berhenti, jika nilai *centroid* baru memiliki nilai yang sama dengan *centroid* lama maka proses iterasi berhenti.

3.4 *Cluster* Optimum

Proses ini dilakukan untuk menentukan *cluster* optimum dengan menggunakan metode *Dunn Index*. *Cluster* dinyatakan optimum apabila nilai *Dunn Index* semakin besar.

3.5 Uji Validitas

Pada tahap ini dilakukan uji validitas untuk melakukan perbandingan antara algoritma *Fuzzy C-Means* serta *K-Means*, dengan menggunakan metode *Partition Coefficient Index* untuk algoritma *Fuzzy C-Means* dan metode *Silhouette Index* untuk algoritma *K-Means* agar mengetahui algoritma mana yang lebih optimum dalam melakukan *clustering*

4. Hasil dan Pembahasan

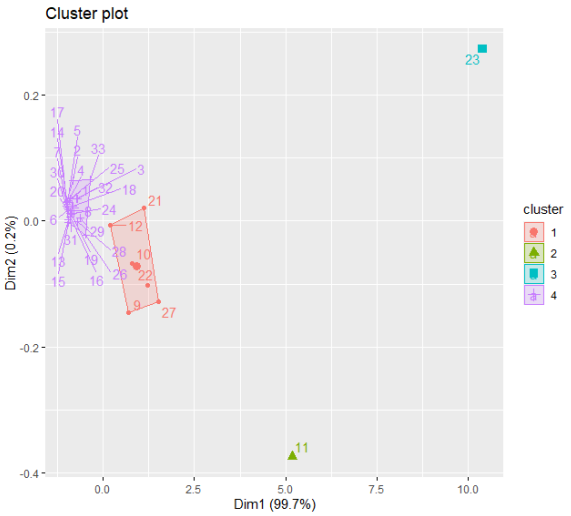
Pada penelitian ini akan menjelaskan hasil yang didapat dari perhitungan yang sudah dilaksanakan. Data tersebut di *cluster* memakai algoritma *Fuzzy C-Means* dan *K-Means*. Selanjutnya mencari *cluster* optimum dari masing-masing algoritma menggunakan metode *Dunn Index*. Lalu dilakukan perbandingan menggunakan uji validitas *Partition Coefficient Index* dan *Silhouette Index*. Data yang digunakan adalah tingkat buta huruf berdasarkan provinsi di Indonesia dari tahun 2015 sampai 2019.

Tabel 1 Data Tingkat Buta Huruf

Provinsi	2015	2016	2017	2018	2019
Aceh	12,21	11,00	6,503	6,503	3,001
Bali	23,83	21,18	16,61	9,136	7,06
Banten	35,27	22,71	25,11	15,54	14,35
Bengkulu	8,27	6,75	5,062	3,375	3,187
Di Yogyakarta	6,59	4,783	7,358	3,311	1,84
...	...	...	...	...	...
Sumatera Utara	66,45	51,57	52,96	36,24	25,09

4.1 Algoritma *Fuzzy C-Means*

Data diproses melalui RStudio yang dikelompokkan memakai algoritma *Fuzzy C-Means* dengan skenario 3 sampai 10 *cluster*. Hasil *output* dari RStudio berupa nilai derajat keanggotaan, pusat *cluster*, jarak setiap data ke pusat *cluster*, data yang masuk dalam setiap *cluster*, jumlah data dalam setiap *cluster*. Perintah RStudio menghasilkan *cluster* dalam bentuk *plot*. **Gambar 3** merupakan contoh *plot* pada 4 *cluster* hasil dari *Fuzzy C-Means* di RStudio.

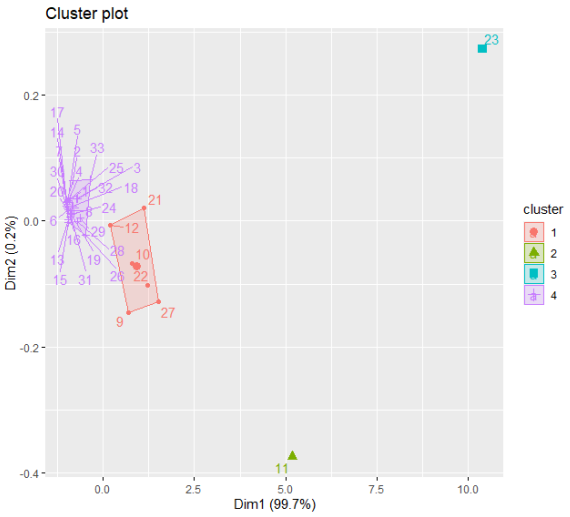


Gambar 3 *Plot* 4 *Cluster* FCM pada Rstudio

Berdasarkan **Gambar 3** *cluster* 1 mempunyai anggota 6 provinsi, *cluster 2* mempunyai anggota 1 provinsi, *cluster* 3 mempunyai anggota 1 provinsi dan *cluster* 4 mempunyai anggota 25 provinsi.

4.2 Algoritma *K-Means*

Data akan diolah menggunakan RStudio yang dikelompokkan menggunakan algoritma *K-Means* dengan skenario 3 sampai 10 *cluster*. Hasil *output* dari RStudio untuk algoritma *K-Means* berupa pusat *centroid*, data yang masuk dalam setiap *cluster*, jumlah data dalam setiap *cluster*. Hasil *cluster* pada RStudio untuk algoritma *K-Means* ditampilkan dalam bentuk *plot*. **Gambar 4** merupakan contoh *plot* pada 5 *cluster* yang dihasilkan dari algoritma *K-Means* di RStudio.



Gambar 4 *Plot* 5 *Cluster K-Means* pada RStudio

Berdasarkan **Gambar 4** *cluster* 1 mempunyai anggota 1 provinsi, *cluster 2* mempunyai anggota 3 provinsi, *cluster* 3 mempunyai anggota 23 provinsi, *cluster* 4 mempunyai anggota 1 provinsi dan *cluster* 5 mempunyai anggota 5 provinsi.

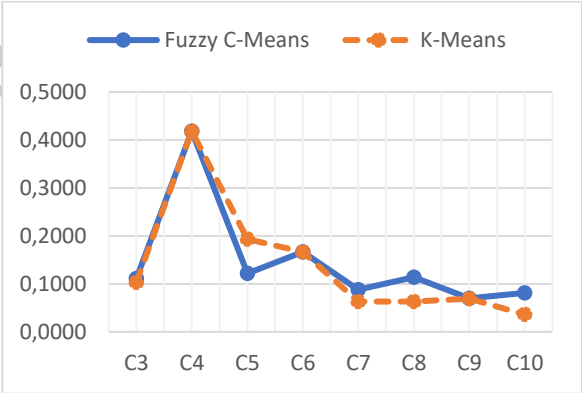
4.3 Penentuan *Cluster* Optimum

Menentukan *cluster* optimum dari algoritma *Fuzzy C-Means* serta *K-Means* dengan menggunakan metode *Dunn Index*. Nilai *Dunn Index* didapat menggunakan persamaan (3.1) Hasil perhitungan *Dunn Index* pada RStudio.

Tabel 2 Nilai *Dunn Index*

C	Fuzzy C-Means	K-Means
3	0,1114	0,1031
4	0,4184	0,4184
5	0,1215	0,1933
6	0,1663	0,1663
7	0,0881	0,0628
8	0,1136	0,0628
9	0,0698	0,0691
10	0,0810	0,0361

Berdasarkan **Tabel 2** *cluster* optimum ditunjukkan dengan semakin besar nilai *Dunn Index*. Pada algoritma *Fuzzy C-Means cluster* optimum ditunjukkan pada 4 *cluster* dengan nilai *Dunn Index* 0,4184. Sedangkan pada algoritma *K-Means cluster* optimum ditunjukkan pada 4 *cluster* dengan nilai *Dunn* Index yang sama seperti algoritma *Fuzzy C-Means* dengan nilai 0,4184. Berikut grafik nilai *Dunn Index* untuk kedua algoritma:



Gambar 5 Grafik *Dunn Index*

4.4 Uji Validitas

4.4.1 Validitas Algoritma *Fuzzy C-Means*

Melakukan uji validitas pada algoritma *Fuzzy C-Means* dengan memakai metode *Partition Coefficient Index* serta menggunakan persamaan (4.1). Berikut hasil perhitungan uji validitas *Partition Coefficient Index* pada RStudio:

Tabel 3 Nilai Validitas *Fuzzy C-Means*

C	Partition Coefficient
3	0,8036
4	0,7795
5	0,6378
6	0,5919
7	0,5097
8	0,4665
9	0,4671
10	0,4507

4.4.2 Validitas Algoritma *K-Means*

Uji validitas pada algorimta *K-Means* menggunakan metode *Silhouette Index*. Nilai *Silhouette Index* didapat dengan menggunakan persamaan (5.1). Berikut hasil perhitungan uji validitas *Silhouette Index* pada RStudio:

Tabel 5 Nilai Validitas *K-Means*

C	Silhouette Index
3	0,7892
4	0,7822
5	0,6549
6	0,5611
7	0,4221
8	0,4495
9	0,4235
10	0,3572

5. Kesimpulan dan Saran

5.1 Kesimpulan

Hasil dari penjabaran dan permasalahan di atas, maka dapat disimpulkan antara lain:

1. Hasil metode *Dunn Index* untuk *cluster* optimum dalam penerapan algoritma *Fuzzy C-Means* adalah 4 *cluster* dengan nilai *Dunn Index* 0,4184 dan untuk algoritma *K-Means* adalah 4 *cluster* dengan nilai *Dunn Index* 0,4184.
2. Berdasarkan hasil uji validitas algoritma *Fuzzy C-Means* serta *K-Means* menunjukkan bahwa algoritma *Fuzzy C-Means* lebih baik dalam melakukan *clustering* dengan nilai validitas 0,8036. Sedangkan pada algoritma *K-Means* memiliki nilai validitas 0,7894. Validitas yang mendekati nilai 1 memiliki kualitas *cluster* yang lebih baik.

5.2 Saran

Berikut beberapa saran dari penelitian ini yaitu:

1. Uji validitas dalam mencari *cluster* optimum, dapat memakai metode selain *Dunn Index*, seperti metode *Elbow*, *Davies Bouldin Index*, *Gap Statistic*, dll.
2. Uji validitas untuk melakukan perbandingan algoritma, dapat menambahkan metode lain, seperti metode *Davies Bouldin Index*, *Calinski Harabasz*, dll.
3. Penelitian ini masih bisa dikembangkan dengan membandingkan berbagai algoritma *clustering* lainnya.

Daftar Pustaka

Berry, A. J. Michael dan Linoff, S. Gordon. 2004. *Data Mining Techniques*. Wiley Publishing, Inc. Indianapolis.

Dey, Debomit. 2019. "Dunn Index and DB Index – Cluster Validity Indices". 5 September 2019.

Everitt, Brian S, Sabine Landau, Morven Lesse, dan Daniel Stahl. 2011. *Cluster Analysis 5th Addition 2011*.

Gelley, Ned dan Jang, Roger. 2000. *Fuzzy Logic Toolbox*. Mathwork, Inc., USA.

Hartigan, John A. *Clustering Algorithm*. Vol. 209. New York: Wiley, 1975.

Izzadin, F. M. 2019. "Optimasi Jumlah Cluster K-Means Dengan Metode Elbow dan Silhouette Untuk Pengelompokan Luas Panen Palawija Kabupaten Magelang Pada Tahun 2017". 20 Juni 2019.

Kusumadewi, S., dan Purnomo, H. 2004. *Aplikasi Logika Fuzzy untuk Pendukung Keputusan*. Yogyakarta: Graha Ilmu.

Prasetyo, E. 2013. Data Mining: *Konsep dan Aplikasi Menggunakan Matlab*. Jakarta: Andi Publisher.