Seminar Nasional Ilmu Komputer (SOLITER)
Vol. 1, Oktober 2017
ISSN: 2597-6303

PERBANDINGAN PENGARUH NILAI *CENTROID* AWAL PADA ALGORITMA K-MEANS DAN K-MEANS++ TERHADAP HASIL *CLUSTER* MENGGUNAKAN METODE *CONFUSION MATRIX*

Muhammad Ni'man Nasir[1], Irwan Budiman[2]
[1,2]Prodi Ilmu Komputer FMIPA ULM
Jl. A. Yani Km 36 Banjarbaru, Kalimantan selatan
Email: nikman.mnn@gmail.com

Abstract

The K-means and K-means ++ algorithms are clustering. The difference between the two is in the selection of the initial centroids. In the K-means the initial centroid selection is randomly selected while the K-means ++ is chosen by probability calculation. This research was conducted to find out how the effect of different initial centroid election from both algorithm to cluster result. Evaluation of cluster results used using confusion matrix method to calculate the accuracy level. From the result of research, it is found that both algorithm have cluster end result can change. This is evidenced from the experiments performed 100 times on both algorithms with iris data sets, the K-means algorithm has 6 different cluster results and the K-means ++ algorithm has 5 different cluster results. K-means ++ algorithm has a higher accuracy than the K-means algorithm. The average accuracy of K-means algorithm is 80.46% and on K-means ++ algorithm is 83.66. The K + Means ++ algorithm has a bigger chance to get the best cluster result, 100 times experiment, K-Means ++ algorithm have 48 experiments while K-Means algorithm has 19 experiments.

Keywords *: k-means, k-means++*, *iris dataset, confusion matrix,* accuracy

Abstrak

Algoritma *K-means* dan *K-means++* merupakan algoritma *clustering*. Perbedaan kedua algoritma tersebut yaitu pada pemilihan *centroid* awal. Pada algoritma *K-means* pemilihan *centroid* awal dipilih secara acak sedangkan algoritma *K-means++* dipilih secara perhitungan probabilitas. Penelitian ini dilakukan untuk mengetahui bagaimana pengaruh dari pemilihan *centroid* awal yang berbeda dari kedua algoritma tersebut terhadap hasil cluster. Evaluasi hasil *cluster* yang digunakan menggunakan metode *confusion matrix* untuk menghitung tingkat akurasi. Dari hasil penelitian didapatkan bahwa kedua algoritma mempunyai hasil akhir *cluster* dapat berubah-ubah. Hal ini dibuktikan dari percobaan yang dilakukan 100 kali pada kedua algoritma dengan *data set iris,* algoritma *K-means* mempunyai 6 hasil *cluster* yang berbeda dan algoritma *K-means++* mempunyai 5 hasil *cluster* berbeda. algoritma *K-means++* Mempunyai rata-rata akurasi yang lebih tinggi dari pada algoritma *K-means.* Rata-rata tingkat akurasi algoritma *K-means* yaitu 80,46 % dan pada algoritma *K-means++* yaitu 83,66. Algoritma K-Means++ mempunyai peluang lebih besar untuk mendapatkan hasil *cluster* terbaik, dari percobaan 100 kali, algoritma K-Means++ terdapat 48 percobaan sedangkan algoritma K-Means terdapat 19 percobaan.

Kata kunci : *k-means, k-means++*, *iris dataset, confusion matrix,* akurasi

ISSN: 2597-6303

1. PENDAHULUAN

Perkembangan teknologi informasi saat ini semakin berkembang pesat, salah satunya pada bidang *Data Mining* (DM) karena besarnya kebutuhan akan nilai tambah dari database skala besar yang makin banyak terakumulasi sejalan dengan pertumbuhan teknologi informasi. Secara umum, *data mining* adalah suatu rangkaian proses untuk menggali nilai tambah berupa ilmu pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data [1].

Data Clustering merupakan salah satu metode *Data Mining* yang bersifat tanpa arahan (*unsupervised*). Ada dua jenis *data clustering* yaitu *hierarchical* (hirarki) *data clustering* dan *non-hierarchical* (non hirarki) *data clustering*. K-Means merupakan salah satu metode *data clustering* non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*/kelompok [2].

K-Means merupakan algoritma *clustering* yang banyak digunakan dalam teknik *clustering* . Algoritma ini dimulai dengan pemilihan secara acak K, yang merupakan banyaknya *cluster* yang ingin dibentuk. Kemudian tetapkan nilai-nilai K secara *random*/acak, untuk sementara nilai tersebut menjadi pusat dari *cluster* atau biasa disebut dengan *centroid*/*mean* [3]. Pemilihan *centroid* awal secara acak ini tidak diketahui sebelumnya apakah dapat berdampak pada hasil akhir klaster. Menentukan titik *centroid* awal secara acak/random dapat ditemukannya beberapa model *cluster*ing yang berbeda [4]. Yang berarti dalam penerapannya algoritma K-Means harus dijalankan beberapa kali sampai mendapatkan hasil kluster yang terbaik.

Algoritma K-Means*++* merupakan pengembangan dari algoritma K-Means yang dikembangkan oleh Arthur dan Vassilvitskii [5]. Pada K-Means++ nilai *centroid* awal tidak dilakukan secara acak seperti pada algoritma K-Means tetapi melalui tahap perhitungan. Setelah nilai *centroid* awal ditentukan maka tahapan selanjutnya sama persis seperti algoritma K-Means. Algoritma K-Means*++* ini belum diketahui apakah juga dapat mempengaruhi hasil kluster seperti pada K-Means yang dapat mempengaruhi hasil *cluster*ing.

Dari perbedaan penentuan nilai *centroid* awal antara algoritma kmeans dan k-mens++ serta pengaruhnya terhadap *hasil cluster*. maka perlu adanya penelitian tentang perbandingan pengaruh nilai *centroid* awal pada algoritma K-Means dan K-Means*++* terhadap hasil *cluster.* Proses pengujian dilakukan menggunakan metode *Confusion Matrix* untuk mengukur tingkat akurasi.

2. METODOLOGI PENELITIAN

2.1 Alat Penelitian

Alat-alat yang digunakan dalam penelitian ini terbagi menjadi dua jenis meliputi perangkat keras dan perangkat lunak.

1) Perangkat Keras
   a) *Processor* AMD A10
   b) RAM 4 GB
   c) Harddisk 1000GB
2) Perangkat Lunak
   a) Windows 8.1 Professional Edition
   b) Webserver Apache
   c) Relational Database Management System, MySQL
   d) PHP 5.6.23
   e) CodeIgniter PHP Framework

2.2 Bahan Penelitian

Bahan yang digunakan dalam penelitian ini adalah *Iris Data Set* yang diperoleh dari situs *https://archive.ics.uci.edu/ml/datasets/Iris.*

2.3 Variabel Penelitian

Variabel yang digunakan dalam penelitian ini meliputi nilai *Centroid* awal, dan *Confusion Matrix.*

2.4. Prosedur Penelitian

Prosedur penelitian yang digunakan pada penelitian ini yaitu : pengumpulan data, *pre-processing,* Eksperimen dan analisis hasil ekperimen. Pada tahap ekperimen dilakukan 100 kali percobaan pada masing-masing algoritma *K-means* dan *K-means++.* Data yang digunakan pada penelitian ini adalah *Iris Data Set*. Variabel yang digunakan dalam penelitian ini meliputi nilai *Centroid* awal, dan *Confusion Matrix.* Adapun prosedur kerja yang akan digunakan dalam penelitian ini adalah sebagai berikut :

2.4.1. Pengumpulan Data

Tahapan ini merupakan tahapan untuk mempersiapkan data yang akan digunakan pada penelitian ini. Data yang digunakan yaitu *Data Set Iris* yang

2.4.2. Pemrosesan Data

Data Preprocessing adalah tahapan untuk menghapus data yang *noise* dan inkonsisten, integrasi antar data yang terkait menjadi satu kesatuan data, pemilihan data

relevan dalam database yang diperlukan dalam keperluan analisis, serta proses transformasi data menjadi bentuk yang siap untuk diproses dalam *data mining*.

2.4.3. Eksperimen

Eksperimen pada tahap ini dilakukan dengan simulasi *clustering* menggunakan program yang telah dibuat. Simulasi pertama yaitu progam melakukan *clustering* menggunakan algoritma K-Means sebanyak percobaan 100 kali dengan nilai *centroid* awal setiap percobaannya dipilih secara acak dan tidak sama pada percobaan sebelumnya. Kemudian simulasi kedua yaitu program melakukan *clustering* menggunakan algoritma K-Means*++* sebanyak percobaan 100 kali dengan nilai *centroid* awal ditentukan berdasarkan perhitungan yang ada pada algoritma K-Means*++.* Dari setiap percobaan tersebut dilakukan evaluasi klustering yaitu nilai *Confusion Matrix* Tahapan pada ekspeimen ini yaitu :

a. *Data mining*

Tahapan proses dalam menemukan suatu pola atau pengetahuan dalam data yang berjumlah besar. Pada penelitian ini data set yang di *cluster* yang nilai Y (*field* kriteria *cluster* sebenarnya) nya dihilangkan agar dapat dibandingkan hasil *cluster*ing yang telah diproses dengan hasil sebenarnya dengan metode *Confusion Matrix*. metode yang digunakan untuk melakukan data mining yaitu *cluster*ing menggunakan algoritma K-Means dan K-Means*++.*

b. *Pattern evaluation*

Tahapan identifikasi pola-pola khusus yang dapat memberikan representasi pengetahuan berdasarkan pengukuran tertentu. Pada tahap ini dilakukan evaluasi dasar terhadap kinerja model *cluster*ing yang telah dihasilkan. Metode yang digunakan untuk evaluasi hasil cluster ini menggunakan metode Confusion Matrix.

Confusion Matrix merupakan salah satu cara untuk melihat kinerja *classifier/supervised learning,* dalam *unsupervised learning* biasa dikenal dengan istilah *matching matrix.* Setiap kolom dari matriks mewakili kelas yang diprediksi, sedangkan setiap baris mewakili kelas yang sebenarnya,.

Jumlah prediksi dalam teknik klasifikasi didasarkan pada jumlah hasil pengujian dengan hasil klasifikasi secara benar atau salah yang telah diprediksi oleh model klasifikasi. Sebagaimana ditunjukkan dalam tabel 10, hasil prediksi dituliskan ke dalam *Confusion Matrix* (tabel kontingensi) di mana nilai kelas sebenarnya disajikan dalam baris tabel, sedangkan kelas hasil prediksi disajikan dalam kolom tabel. *Confusion Matrix* menunjukkan pola prediksi *classifier* terhadap masing-masing kelas [6].

Tabel 1. Confusion Matrix Dua Kelas

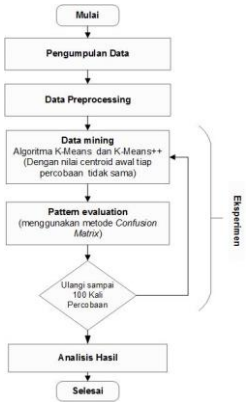
		Predicted	
		Yes	No
Actual	Yes	TP	FN
	No	FP	TN

Keterangan:
TP= *True Positives* FP= *False Positives*
TN= *True Negatives* FN= *False Negatives*

True Positives menunjukkan jumlah kelas positif yang tepat diprediksi sebagai kelas positif oleh model. *True Negative* menunjukkan jumlah kelas negatif yang tepat diprediksi sebagai kelas negatif oleh model. *False Positives* menunjukkan jumlah kelas negatif yang keliru diprediksi sebagai kelas positif oleh model. *False Negatives* menunjukkan jumlah kelas positif yang keliru diprediksi sebagai kelas negatif oleh model [6].

2.1.3. Analisis Hasil Eksperimen

Eksperimen yang telah dilakukan kemudian dianalisis. Analisis pada eksperimen ini yaitu membandingkan hasil evaluasi klustering berupa nilai Confusion Matrix dari 100 percobaan yang telah dilakukan pada tiap percobaan.



Gambar 1. Alur prosedur penelitian

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah *Iris Data Set* yang diperoleh dari *https://archive.ics.uci.edu/ml/datasets/Iris*.
Data set iris ini mempunyai jumlah data sebanyak 150 dengan 3 kategori yaitu 50 Iris virginica, 50 Iris setosa, dan 50 Iris Versicolor. Jumlah Atribut pada data set iris ini ada 5 yaitu 4 atribut bertipe bilangan, dan 1 atribut bertipe kategori. 5 atribut tersebut adalah Sepal length (cm), Sepal width (cm), Petal length(cm) , Petal width (cm), Species (Iris-setosa, Iris-virginica dan Iris-versicolor).

3.2. *Pre-processing*

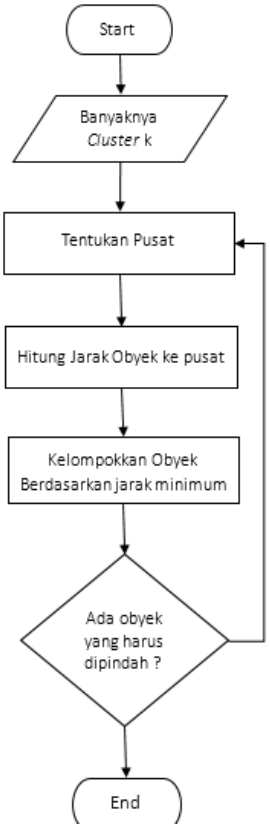
Tahap *preprocessing* dalam penelitian ini hanya menggunakan data cleaning agar data dapat diolah pada tahap data mining. *Data cleaning* adalah sebuah proses dalam metode kdd yang bertujuan untuk menghilangkan *noise* data yang tidak konsisten Pada tahap data cleaning ini yaitu dengan menghilangkan atribut species karena tidak digunakan pada tahap data mining.

3.3. Eksperimen

Tahap ini dilakukan dengan simulasi *clustering*. Simulasi pertama yaitu melakukan *clustering* menggunakan algoritma K-Means sebanyak 100 kali percobaan dengan nilai *centroid* awal setiap percobaannya dipilih secara acak dan tidak sama pada percobaan sebelumnya. Kemudian simulasi kedua yaitu melakukan *clustering* menggunakan algoritma K-Means*++* sebanyak 100 kali percobaan dengan nilai *centroid* awal ditentukan berdasarkan perhitungan yang ada pada algoritma K-Means*++.* Dari setiap percobaan tersebut dilakukan evaluasi *clustering* yaitu dengan metode *Confusion Matrix.*

3.3.1. Algoritma K-Means

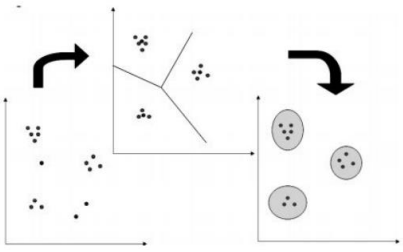
Algoritma K-Means merupakan algoritma *clustering* yang mengelompokkan data berdasarkan titik pusat *cluster* (*centroid*) terdekat dengan data. Pada algoritma K-Means, *centroid* awal dipilih secara *random* oleh komputer. Ukuran kemiripan yang digunakan dalam *cluster* adalah fungsi jarak. Sehingga pemaksimalan kemiripan data didapatkan berdasarkan jarak terpendek antara data terhadap titik *centroid* [7].



Gambar 2. Diagram alir algoritma K-Means

Data *cluster*ing menggunakan algoritma K-Means ini secara umum dilakukan dengan algoritma dasar sebagai berikut :

(1) Tentukan jumlah *cluster*
(2) Alokasikan data ke dalam *cluster* secara random
(3) Hitung *centroid*/rata-rata dari data yang ada di masing-masing *cluster*
(4) Alokasikan masing-masing data ke *centroid*/rata-rata terdekat
(5) Kembali ke Step 3, apabila masih ada data yang berpindah *cluster* atau apabila perubahan nilai *centroid*, ada yang di atas nilai *threshold* yang ditentukan atau apabila perubahan nilai pada *objective function* yang digunakan di atas nilai *threshold* yang ditentukan.



Gambar 3. Ilustrasi Algoritma K-Means

ISSN: 2597-6303

Pada penelitian ini algoritma K-Means akan diuji coba dengan mengulang proses algoritma K-Means sebanyak 100 kali kemudian hasil *cluster* dicatat untuk di evaluasi. Hasil percobaan dapat dilihat pada tabel 1.

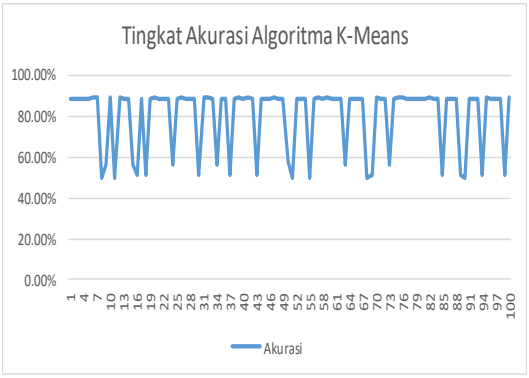
Tabel 2. Evaluasi hasil *cluster* seluruh percobaan algoritma K-Means

Percobaan ke-	***Confusion Matrix* (%)**
1	88.6667
2	88.6667
3	89.3333
4	56.0000
5	88.6667
6	88.6667
7	88.6667
9	89.3333
8	89.3333
10	50.0000
11	50.0000
12	89.3333
13	88.6667
14	88.6667
15	56.0000
16	51.3333
18	51.3333
19	88.6667
20	89.3333
21	88.6667
22	88.6667
23	88.6667
24	56.0000
25	88.6667
26	89.3333
27	88.6667
28	88.6667
29	88.6667
30	51.3333
31	89.3333
32	89.3333
33	88.6667
34	56.0000
35	88.6667
36	88.6667
37	51.3333
38	88.6667
39	89.3333
40	88.6667
41	89.3333
42	88.6667
43	51.3333
44	88.6667
45	88.6667
46	88.6667
47	89.3333
48	88.6667
49	88.6667
50	57.3333
51	50.0000
52	88.6667
53	88.6667
54	88.6667
55	50.0000
56	88.6667
57	89.3333
58	88.6667
59	89.3333
60	88.6667
61	88.6667
62	88.6667
63	56.0000
64	88.6667
65	88.6667
66	88.6667
67	88.6667
68	50.0000
69	51.3333
70	89.3333
71	88.6667
72	88.6667
73	56.0000
74	88.6667
75	89.3333
76	89.3333
77	88.6667
78	88.6667
79	88.6667
80	88.6667
81	88.6667
82	89.3333
83	88.6667
84	88.6667
85	51.3333
86	88.6667
87	88.6667
88	88.6667
89	51.3333
90	50.0000
91	88.6667
92	88.6667
93	88.6667
94	51.3333
95	89.3333
96	88.6667
97	88.6667
98	88.6667
99	51.3333
100	89.3333

ISSN: 2597-6303

Pada Percobaan algoritma K-Means ini ketika algoritma tersebut dilakukan sebanyak 100 kali, nilai evaluasi *cluster* nilai confussion martix mengalami perubahan nilai. Percobaan pertama dengan percobaan kedua mempunyai hasil yang sama, tetapi ketika algoritma dijalan kan kembali pada percobaan ketiga, hasil *cluster* yang didapatkan tidak sama dengan hasil *cluster* percobaan sebelumnya. Begitu pula pada percobaan selanjutnya, hasil *cluster* dapat berubah-ubah tetapi juga terkadang beberapa percobaan mempunyai hasil yang sama dengan percobaan sebelumnya. Dari 100 percobaan algoritma K-Means ini terdapat 6 hasil *cluster* yang berbeda.

Dari grafik pada gambar 2, yang merupakan grafik nilai confusion matrik pada percobaan algoritma K-Means mempunyai garis grafik yang tidak lurus, hal ini dikarenakan hasil *cluster* tidak selalu sama ketika algoritma K-Means dijalankan beberapa kali. Berdasarkan percobaan yang telah dilakukan sebanyak 100 kali pada algoritma k-mean dengan *centroid* awal yang berbeda didapatkan hasil *cluster* yang berubah-ubah, hal ini dapat disimpulkan bahwa *centroid* awal yang dipilih secara acak/random pada algoritma K-Means dapat mempengaruhi hasil *cluster*.



Gambar 4. Grafik tingkat akurasi pada percobaan algoritma K-Means

Tabel 3. Hasil *cluster* berbeda pada percobaan Algoritma K-Means

No	Jumlah Percobaan	Tingkat Akurasi
1	19	89.3333 %
2	58	88.6667 %
3	10	51,3333 %
4	6	56%
5	1	57%
6	6	50%

Hasil *cluster* terbaik pada percobaan algoritma K-Means terdapat tingkat akurasi sebesar 89.3333 % dengan jumlah percobaan sebanyak 19. Hasil *cluster* terbaik kedua terdapat pada hasil yang mempunyai tingkat akurasi 88,6667% dengan jumlah percobaan sebanyak 58. Sedangkan untuk hasil *cluster* dengan tingkat akurasi dibawah 60% merupakan hasil *cluster* yang kurang baik, hasil *cluster* kurang baik ini ada 4 hasil dengan jumlah percobaan sebanyak 23. Pada percobaan algoritma K-Means ini, hasil *cluster* yang paling banyak muncul adalah hasil cluter terbaik kedua yaitu hasil *cluster* yang tingkat akurasi sebesar 88.6667 %.

3.3.2. Algoritma K-Means++

Pada algoritma K-Means *centroid* awal dipilih secara acak dari kumpulan data. Meskipun pendekatan ini sederhana dan cepat, akan tetapi terkadang menghasilkan hasil yang jauh dari optimal, karena tidak ada jaminan akurasi. Beberapa penelitian telah dilakukan untuk memperbaiki kekurangan algoritma K-Means, salah satunya yaitu algoritma K-Means*++.* Pada K-Means++ *centroid* awal dipilih dengan probabilitas tertentu . probabilitas pemilihan titik sebagai *centroid* sebanding dengan jarak *centroid* terdekat yang sudah dipilih [5].

Berikut adalah algoritma K-Means*++* :

(1) Pilih pusat awal c_1 seragam secara acak dari X. Hitung vektor yang berisi *square distance* antara semua titik dalam dataset dan c_1,
(2) Pilih pusat kedua c_2 dari X secara acak dari distribusi probabilitas
(3) Hitung ulang jarak vektor
(4) Pilih pusat c_1 berturut-turut dan menghitung ulang jarak vektor.
(5) Jika k *center* telah dipilih, langkah selanjutnya standar algoritma K-Means.

Pada penelitian ini algoritma K-Means++ akan diuji coba dengan mengulang proses algoritma K-Means sebanyak 100 kali kemudian hasil *cluster* dicatat untuk dievaluasi. Hasil percobaan dapat dilihat pada tabel 3.

ISSN: 2597-6303

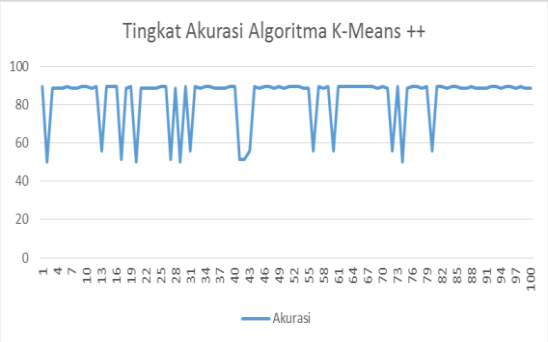
Tabel 4. Hasil *cluster* berbeda pada percobaan Algoritma K-Means++

Percobaan ke-	***Confusion Matrix* (%)**
1	89.33333
2	50
3	88.66667
4	88.66667
5	88.66667
6	89.33333
7	88.66667
8	88.66667
9	89.33333
10	89.33333
11	88.66667
12	89.33333
13	56
14	89.33333
15	89.33333
16	89.33333
17	51.33333
18	88.66667
19	89.33333
20	50
21	88.66667
22	88.66667
23	88.66667
24	88.66667
25	89.33333
26	89.33333
27	51.33333
28	88.66667
29	50
30	89.33333
31	56
32	89.33333
33	88.66667
34	89.33333
35	89.33333
36	88.66667
37	88.66667
38	88.66667
39	89.33333
40	89.33333
41	51.33333
42	51.33333
43	56
44	89.33333
45	88.66667
46	89.33333
47	89.33333
48	88.66667
49	89.33333
50	88.66667
51	89.33333
52	89.33333
53	89.33333
54	88.66667
55	88.66667
56	56
57	89.33333
58	88.66667
59	89.33333
60	56
61	89.33333
62	89.33333
63	89.33333
64	89.33333
65	89.33333
66	89.33333
67	89.33333
68	89.33333
69	88.66667
70	89.33333
71	88.66667
72	56
73	89.33333
74	50
75	88.66667
76	89.33333
77	89.33333
78	88.66667
79	89.33333
80	56
81	89.33333
82	89.33333
83	88.66667
84	89.33333
85	89.33333
86	88.66667
87	88.66667
88	89.33333
89	88.66667
90	88.66667
91	88.66667
92	89.33333
93	89.33333
94	88.66667
95	89.33333
96	89.33333
97	88.66667
98	89.33333
99	88.66667
100	88.66667

Tabel 3. Lanjutan Hasil *cluster* berbeda pada percobaan Algoritma K-Means++

Pada Percobaan algoritma K-Means++ ini ketika algoritma tersebut dilakukan sebanyak

100 kali, nilai evaluasi *cluster* yaitu nilai confussion martix mengalami perubahan nilai. Hasil percobaan satu dengan percobaan lainnya terkadang mempunyai nilai yang sama tetapi juga mempunyai nilai yang berbeda. Dari 100 kali percobaan terdapat sebanyak 5 hasil *cluster* yang berbeda.



Gambar 5. Grafik tingkat akurasi pada percobaan algoritma K-Means++

Tingkat akurasi pada percobaan algoritma K-Means++ mempunyai garis grafik yang tidak lurus, hal ini dikarenakan hasil *cluster* tidak selalu sama ketika algoritma K-Means dijalankan beberapa kali. Berdasarkan percobaan yang telah dilakukan sebanyak 100 kali pada algoritma K-Means++ dengan hasil *cluster* yang didapatkan berubah-ubah,. hal ini dapat disimpulkan bahwa *centroid* awal yang dipilih secara perhitungan ada algoritma K-Means++ juga dapat mempengaruhi hasil *cluster* sama seperti algoritma K-Means yang menentukan *centroid* awalnya secara acak/random.

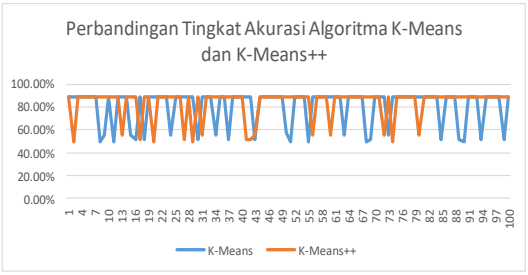
Tabel 5. Hasil *cluster* berbeda pada percobaan Algoritma K-Means++

No	Jumlah Percobaan	*Tingkat Akurasi*
1	49	89.3333 %
2	36	88.6667 %
3	4	51,3333 %
4	7	56%
5	4	50%

Hasil *cluster* terbaik pada percobaan algoritma K-Means++ terdapat pada hasil yang mempunyai tingkat akurasi sebesar 89.3333% dengan jumlah percobaan sebanyak 49. Hasil *cluster* terbaik kedua terdapat pada hasil yang mempunyai tingkat akurasi 88,6667% dengan jumlah percobaan sebanyak 36. Sedangkan untuk hasil *cluster* dengan tingkat akurasi dibawah 60% merupakan hasil *cluster* yang kurang baik, hasil *cluster* kurang baik ini ada 3 hasil dengan jumlah percobaan sebanyak 15. Pada percobaan algoritma K-Means ini, hasil *cluster* yang paling banyak muncul adalah hasil cluter terbaik pertama yaitu hasil *cluster* yang mempunyai tingkat akurasi sebesar 89.3333%.

Hasil akhir cluster ditentukan dari nilai centroid awal yang paling dekat dengan salah satu centroid akhir cluster dari hasil cluster tersebut. Misalkan pada percobaan 1 algoritma K-Means++ yang mempunyai nilai akurasi 89,33%, jika centroid awal pada percobaan tersebut diukur jarak dengan centroid akhir dari masing-masing hasil cluster tersebut. Maka didapatkan rata-rata jarak yang paling kecil merupakan jarak antara centroid awal dengan centroid akhir hasil cluster yang mempunyai akurasi 89,33%.

3.4. Analisis Hasil



Gambar 6. Grafik Perbandingan Tingkat Akurasi K-Means dan K-Means++

Dari grafik pada gambar 6, yang merupakan grafik perbandingan tingkat akurasi pada algoritma K-Means dan K-Means++ pada grafik gambar 10, algoritma K-Means++ juga cenderung lebih stabil dari pada algoritma K-Means.

Tabel 34. Rata-rata jarak centroid awal percobaan 1 algoritma k-means++ dengan centroid akhir masing-masing hasil cluster

Akurasi Hasil Cluster	Rata-rata jarak dari centroid awal pada percobaan 1 K-Means++
89,33%	1.95107
88,67%	1.960296
56,00%	3.280364
51,33%	3.907888
50,00%	3.30738

Hasil akhir cluster ditentukan dari nilai centroid awal yang paling dekat dengan salah

satu centroid akhir cluster dari hasil cluster tersebut. Misalkan pada percobaan 1 algoritma K-Means++ yang mempunyai nilai akurasi 89,33%, jika centroid awal pada percobaan tersebut diukur jarak dengan centroid akhir dari masing-masing hasil cluster tersebut. Maka didapatkan rata-rata jarak yang paling kecil merupakan jarak antara centroid awal dengan centroid akhir hasil cluster yang mempunyai akurasi 89,33%.

Pengaruh dari centroid awal pada algoritma K-Means yaitu hasil akhir cluster dapat berubah-ubah sesuai dengan centroid awal yang dipilih. Dari percobaan yang dilakukan 100 kali terdapat 6 hasil cluster yang berbeda. Pengaruh dari centroid awal yang ditentukan secara perhitungan pada algoritma K-Means++ yaitu hasil akhir cluster juga dapat berubah-ubah sesuai dengan centroid awal yang dipilih. Dari percobaan yang dilakukan 100 kali terdapat 5 hasil cluster yang berbeda. Hasil *cluster* pada algoritma K-Means dan K-Means mempunyai hasil *cluster* terbaik yang sam yaitu hasil *cluster* yang mempunyai tingkat akurasi sebesar 89.3333 %, Rata-rata tingkat akurasi algoritma K-Means++ lebih tinggi yaitu 83,66 % sedangkan algoritma K-Means yaitu 80,46. Algoritma K-Means++ mempunyai peluang lebih besar untuk mendapatkan hasil *cluster* terbaik yaitu algoritma K-Means++ terdapat 48 percobaan yang merupakan hasil *cluster* terbaik. sedangkan pada algoritma K-Means hanya terdapat 19 percobaan saja.

Berdasarkan dari hasil percobaan yang dilakukan sebanyak 100 kali pada masing-masing algoritma yaitu algoritma K-Means yang *centroid* awal dipilih secara random maupun K-Means++ yang *centroid* awal dipilih melalui perhitungan, *centroid* awal kedua algoritma tersebut sama-sama mempengaruhi hasil *cluster*. hal ini dibuktikan dengan nilai akurasi yang berubah-ubah. Dilihat dari jumlah hasil *cluster*, algoritma K-Means++ lebih unggul dibandingkan dengan algoritma K-Means, pada algoritma K-Means++ hasil *cluster* hanya terdapat 5 hasil *cluster* yang berbeda sedangkan algoritma K-Means terdapat 6 hasil *cluster* yang berbeda.hal ini membuktikan bahwa *centroid* awal dengan pemilihan secara random lebih mempengaruhi banyaknya hasil *cluster* berbeda yang dihasilkan dari pada *centroid* awal yang dipilih dengan perhitungan.

Selain unggul pada jumlah hasil *cluster*, algoritma K-Means++ juga lebih unggul pada peluang kemunculan hasil *cluster* terbaik. hasil *cluster* terbaik pada peng*cluster*an data set iris adalah hasil *cluster* yang mempunyai tingkat akurasi sebesar 89.3333 %. pada algoritma K-Means++ dari 100 kali percobaan terdapat 48 percobaan yang merupakan hasil *cluster* terbaik. sedangkan pada algoritma K-Means hanya terdapat 19 percobaan saja. Sehingga dapat dikatakan bahwa algoritma K-Means++ dengan *centroid* awal yang dipilih secara perhitungan mempunyai kemungkinan lebih besar untuk mendapatkan hasil *cluster* terbaik dari pada algoritma K-Means.

4. SIMPULAN

Simpulan yang dapat diambil pada penelitian ini adalah :

a. Pengaruh dari centroid awal pada algoritma K-Means yaitu hasil akhir cluster dapat berubah-ubah sesuai dengan centroid awal yang dipilih. Dari percobaan yang dilakukan 100 kali terdapat 6 hasil cluster yang berbeda.
b. Pengaruh dari centroid awal yang ditentukan secara perhitungan pada algoritma K-Means++ yaitu hasil akhir cluster juga dapat berubah-ubah sesuai dengan centroid awal yang dipilih. Dari percobaan yang dilakukan 100 kali terdapat 5 hasil cluster yang berbeda.
c. algoritma *K-means++* Mempunyai rata-rata akurasi yang lebih tinggi dari pada algoritma *K-means.* Rata-rata tingkat akurasi algoritma *K-means* yaitu 80,46 % dan pada algoritma *K-means++* yaitu 83,66. Algoritma K-Means++ mempunyai peluang lebih besar untuk mendapatkan hasil *cluster* terbaik, dari percobaan 100 kali, algoritma K-Means++ terdapat 48 percobaan sedangkan algoritma K-Means terdapat 19 percobaan.

DAFTAR PUSTAKA

[1] Nasir, M. N., "**Perbandingan Pengaruh Nilai Centroid Awal Pada Algoritma K-Means Dan K-Means++ Terhadap Hasil Cluster Menggunakan Metode Confusion Matri dan Sillhouette Coefficient"**. Skripsi Program Studi Ilmu Komputer, Universitas Lambung Mangkurat, Banjarbaru.

[2] Asroni & R. Adrian, **"Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang "**. Jurnal

Ilmiah Semesta Teknika. 18(1):76-82. 2015

[3] Bouveyron, C., & Brunet-Saumard, **"Model-based clustering of high-dimensional data: A review"**. Computational Statistics & Data Analysis, *71*, 52-78, 2014.

[4] Agusta, Yudi, "**K-Means Penerapan, Permasalahan dan Metode Terkait**". Jurnal Sistem dan Informatika, 3 : 47-60, 2013.

[5] Arthur, D. & Vassilvitskii, "**K-Means ++ : The Advantages of Careful Seeding. Proceedings of ACM-SIAM Symposium on Discrete Algorithms"**. 8 : 1–11, 2007.

[6] Omary, Z., **"Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised Learning"**. International Journal for Infonomics (IJI). 3(3) : 314-325. 2010

[7] Susanto, E.B., **"Evaluasi Hasil Klaster Pada Dataset Iris, Soybean-small, Wine Menggunakan Algoritma Fuzzy C-Means dan Kmeans++".** Surya Informatika. 2(1) : 6-13, 2016.