ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 9, Nomor 2, Juli 2020

OPTIMASI ALGORITMA KLASTER DINAMIS PADA K-MEANS DALAM PENGELOMPOKAN KINERJA AKADEMIK MAHASISWA (STUDI KASUS: UNIVERSITAS PENDIDIKAN GANESHA)

Komang Ariasa[1], I Gede Aris Gunadi[2], I Made Candiasa[3]

[1,2,3]Program Studi Ilmu Komputer, Universitas Pendidikan Ganesha, Singaraja, Indonesia

e-mail: komangariasa@undiksha.ac.id[1], igedearisgunadi@undiksha.ac.id[2], candiasa@undiksha.ac.id[3]

Abstrak

Penelitian ini difokuskan untuk perbaikan algoritma klaster dinamis pada *k-means* menggunakan inisiasi *centroid* awal berbasis metode *mean*. Data penelitian menggunakan kinerja akademik 765 orang berasal dari 38 prodi Undiksha, perhitungan klastering berdasarkan nilai UN, rapor dan perkembangan kinerja akademik mahasiswa selama 6 semester. Perbandingan algoritma terbaik diuji tingkat validitasnya menggunakan metode *Cluster Variance* (V), *Davies Bound Index* (DBI), *Partition Coefficient* (PC) dan *Sum Squared Error* (SSE) pada algoritma *k-means* tradisional, *k-means* dinamis dan *k-means* dinamis berbasis *mean*. Berdasarkan pengujian diperoleh 5 jumlah klaster ideal pada metode *k-means* dinamis berbasis inisiasi *centroid*, dengan nilai terbaik PC 0,20176, SSE 2,15152, *variance* terkecil 0,259281 dan DBI 0,168236. Secara keseluruhan optimasi algoritma *k-means* dinamis berbasis *mean* menghasilkan rata-rata kualitas klaster yang lebih baik dan jumlah interasi yang konstan pada setiap pengujian dibanding algoritma k-means lain dalam evaluasi PC, SSE, dan *cluster variance*. Hasil pengujian dapat digunakan sebagai salah satu metode terbaik dalam evaluasi kinerja akademik mahasiswa serta acuan pengambilan keputusan dalam menentukan kebijakan akademik universitas.

Kata kunci: *k-means*, klastering, klaster dinamis, *mean based*

Abstract

This study was focused on the improvement of dynamic clustering algorithm on k-means using the initial centroid initiation based on the mean method. The data was collected by looking at the academic performance of 765 peoples from 38 study programs in Ganesha University of Education (Undiksha), the clustering calculation was based on national examination result, the progress report and the progress of academic performance of college student during six semesters. The best algorithm comparison was tested related to its validity using some methods namely Cluster Variance (V), Davies Bound Index (DBI), Partition Coefficient (PC) and Sum Squared Error (SSE) on algorithm traditional k-means, dynamic k-means and dynamic k-means based mean. The result showed that there were five ideal clusters on k-means dynamic based on centroid initiation method, with the best value was PC 0,20176, SSE 2,15152, the lowest variance was 0,259281 and DBI 0,168236. Generally, the dynamic k-means algorithm based mean optimization results the better cluster quality and the constant number of interaction in every testing compared with the other k-means algorithm in PC, SSE, and cluster variance evaluations. This result can be used as one of best method in evaluating the college student's academic performance as well as the reference in decision making especially in determining university academic policy.

Keywords : *k-means, clustering, dynamic k-means, mean based*

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 9, Nomor 2, Juli 2020

PENDAHULUAN

Dynamic k-means clustering merupakan pengembangan algoritma *k-means* untuk mengecek ulang kualitas *cluster* pada setiap iterasi, memungkinkan terjadinya perubahan dalam jumlah *cluster* untuk memenuhi keabsahan kualitas *cluster*. Algoritma klaster dinamis bertujuan untuk meningkatkan kualitas *cluster* sehingga menghasilkan angka yang optimal dari *cluster* [1]. Metode ini merupakan pengembangan dari algoritma *k-means* tradisional, telah dilakukan sebelumnya oleh (Ahamed & Hareesha, 2012) dan (Widiarini & Wahono, 2015) dengan mengusulkan algoritma klaster dinamis dalam menetapkan jumlah *cluster* (k) agar dapat menghasilkan kualitas *cluster* yang optimal, sehingga dapat memberikan hasil pemetaan potensial lebih baik dan tepat [1]. Namun penelitian pengembangan ini masih memiliki kekurangan terkait komputasi dalam pemrosesan data. Algoritma *dynamic k-means* memiliki kemampuan untuk mencari jumlah *cluster* ideal, namun terdapat kekurangan dalam penentuan titik *centroid* (pusat *cluster*) yang masih dipilih secara acak [2]. Sehingga kesalahan penentuan *centroid* awal akan mempengaruhi jumlah proses iterasi dan waktu komputasi.

Sebagai salah satu perguruan tinggi terbesar di Bali Utara, Undiksha sejak tahun 2012 telah melakukan penghimpun data dalam jumlah besar melalui sistem informasi yang digunakan. Jumlah data yang banyak ini membuka peluang untuk dihasilkan informasi yang berguna bagi pihak universitas [3] dan menjadi modal cukup penting untuk memperoleh pengetahuan-pengetahuan yang dapat menjawab pertanyaan pengelola perguruan tinggi terkait permasalahan yang dihadapi [4]. Dalam kasus segmentasi, teknik *clustering* dapat digunakan untuk mengevaluasi kinerja akademik melalui berbagai atribut yang mendukung prestasi akademik mahasiswa. Dengan adanya jumlah data kinerja akademik universitas yang dihimpun semakin berkembang dan memiliki karakteristik data yang beragam, maka diperlukan suatu metode pengelompokan data yang lebih baik dari *k-means* biasa. Dengan berbagai kelebihan pada klaster dinamis dalam menentukan jumlah klaster ideal, maka metode klastering *k-means* dikolaborasikan dengan metode pendekatan untuk penentuan pusat awal *cluster* yang diharapkan dapat mengurangi waktu komputasi untuk *dataset* yang besar. Metode yang akan diterapkan dalam penelitian adalah inisiasi *centroid* menggunakan metode *mean,* dimana hasil metode rata-rata *centroid* akan digunakan untuk inisialisasi awal pada perhitungan metode *k-means* dan dilanjutkan evaluasi klaster. Hasil penentuan titik *centroid* ini akan berpengaruh pada proses dalam menghasilkan *cluster* konvergen. Kinerja dari algoritma klastering dibandingkan dalam hal kemurnian, menormalkan informasi timbal balik dan waktu yang diambil untuk membentuk sebuah *cluster*. Data kinerja akademik yang dikumpulkan dikelompokkan sesuai dengan karakteristik yang serupa dan membentuk klaster [5].

Penelitian ini difokuskan untuk penerapan perbaikan algoritma *k-means* menggunakan klaster dinamis dengan inisiasi *centroid* awal berbasis metode *mean*. Pengujian algoritma menggunakan data kinerja akademik yang telah dikumpulkan, kemudian dikelompokkan sesuai dengan karakteristik yang serupa dan membentuk klaster ideal. Melalui implementasi ini akan diketahui algoritma terbaik antara *k-means* tradisional, *k-means* dinamis, *k-means* dinamis dengan inisiasi *centroid* serta diharapkan dapat menjadi acuan algoritma terbaik dalam klasterisasi kinerja akademik mahasiswa.

METODE

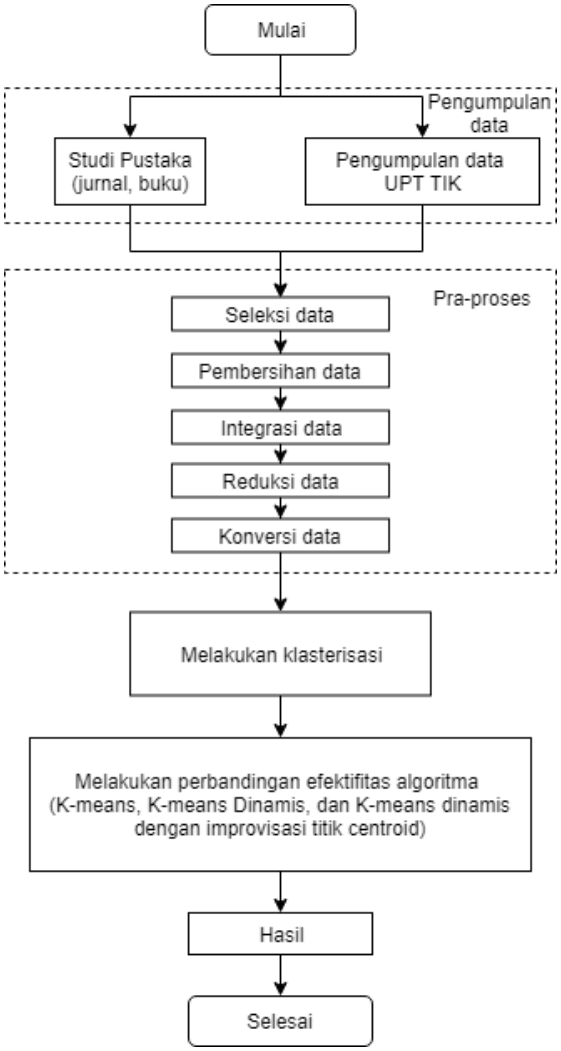
A. Pengumpulan Data

Data utama bersumber dari data akademik dan riwayat mahasiswa yang

tersimpan dalam *database* UPT TIK Undiksha. Data yang digunakan adalah beberapa *dataset* yang didapatkan dari hasil penambangan Sistem Informasi Akademik dan Sistem Penerimaan Mahasiswa Baru Undiksha, dimana data latih dan data uji merupakan mahasiswa angkatan 2017 dengan perkembangan kinerja akademik hingga 6 semester dengan jumlah 765 orang dan memiliki latar belakang yang beragam. Data yang diperlukan yaitu data pribadi mahasiswa, riwayat studi selama kuliah, riwayat pendidikan saat SMA/SMK/MA, dan keadaan ekonomi mahasiswa.

B. Pra Proses

Variabel yang akan dijadikan acuan dalam proses klastering adalah nilai ujian nasional, nilai rapor, dan IPK. Proses yang terjadi pada tahap ini mencakup empat hal yaitu pembersihan data, integrasi data, reduksi data dan konversi data. Gambar 1 merupakan alur kerja yang diterapkan pada penelitian.

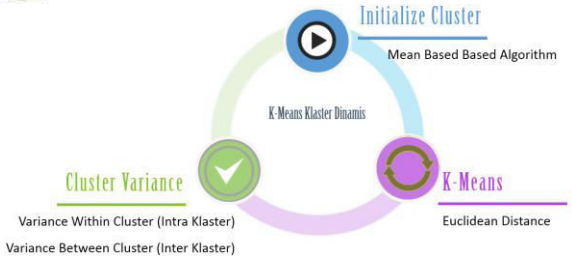


Gambar 1. Rancangan alur penelitian

C. Klastering Dengan *K-means* Dinamis Berbasis *Mean*

Secara garis besar tahap pengembangan algoritma klaster dinamis dengan inisiasi *centroid* awal menggunakan metode *mean* terdiri dari tiga proses utama yaitu inisialisasi klaster awal, proses *clustering* menggunakan algoritma *k-means*, dan perhitungan *cluster variance* meliputi intra dan inter klaster [6]. Proses utama implementasi algoritma secara garis besar dapat dilihat pada Gambar 2.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 9, Nomor 2, Juli 2020



Gambar 2. Algoritma klaster dinamis pada *k-means* dengan *mean based*

Gambar 3 menjelaskan alur penggabungan dari kedua algoritma. Titik *centroid* menggunakan teknik pencarian data dengan metode rata-rata (*mean)*. Inisialisasi dimulai dari proses perhitungan *euclidean distance* pada setiap data $(x, 0)$. Kemudian mengurutkan data berdasarkan jarak pusat klaster yang dihasilkan pada setiap data. Pengurutan bisa dilakukan dari kecil ke besar atau sebaliknya. Data yang sudah diurutkan dibagi kedalam klaster yang diinginkan, dalam hal ini berjumlah tiga klaster. Setiap data yang terlibat harus masuk kedalam salah satu klaster yang tersedia. Selanjutnya adalah menghitung titik pusat klaster (*centroid*) baru dengan mencari rata-rata disetiap klaster. Hasil *centroid* sebelumnya akan digunakan sebagai *centroid* awal pada tiga klaster yang disediakan sebelumnya, kemudian akan dilanjutkan dengan perhitungan menggunakan algoritma *k-means* dinamis.

Algoritma klaster dinamis memiliki cara kerja diawal sama dengan algoritma *k-means,* namun pada tahap akhir terdapat validasi klaster yaitu jika jarak *intra* lebih kecil dan jika jarak *intra* lebih besar, maka algoritma menghitung klaster baru dengan menambahkan *counter* k dengan satu atau k=k+1 disetiap iterasi sampai memenuhi batas validitas kualitas *cluster* yang berkualitas [7].

Istilah *inter* adalah minimum jarak antar pusat *cluster*, *inter* digunakan untuk mengukur pemisahan antar *cluster* yang didefinisikan [7]:

$$inter = min||mk - mkk||\forall k = 1,2,\ldots, k-1\ dan\ k = k+1,\ldots,K \quad (1)$$

Dimana mk adalah jarak pusat klaster sebelumnya, mkk adalah jarak pusat klaster berikutnya, dan k adalah klaster.

Untuk mengukur kekompakan dari suatu kelompok menggunakan *intra.* Deviasi digunakan untuk memeriksa kedekatan titik data setiap *cluster* [7], dan dihitung menggunakan:

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Xi - Xm)^2} \quad (2)$$

Dimana n merupakan jumlah data, Xi merupakan data, dan Xm merupakan *centroid*. *Pseudocode* algoritma dapat dilihat pada Tabel 1 dan Tabel 2.

Tabel 1. *Pseudocode* algoritma inisiasi *centroid* berbasis *mean*

Input:
k: jumlah klaster, inisiasi jumlah klaster awal adalah k = 2
mean: nilai *centroid* awal menggunakan metode *mean*
D: {d_1, d_2,…..,d_n} data set berisi n objek
changed: *boolean*

Output:
c: {c_1,c_2,….,c_k} nilai *centroid* klaster
L: {*l*(e) | e = 1,2,….,n} anggota klaster dari D

Method:
1. Membuat partisi sejumlah k dari segmentasi yang akan dibentuk.
2. Lakukan inisiasi titik pusat klaster menggunakan algoritma *mean*.
3. Hitung e*uclidean distance* menggunakan persamaan

$$d = \sum_{i=1}^{k}\sum_{x_\lambda \epsilon classi}^{n}\sqrt{\sum_{j=1}^{q}(x_{\lambda,j} - x_{i,j})^2} \text{ dan} \quad (3)$$

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2} \quad (4)$$

4. Lakukan pengurutan data dari terbesar ke terkecil.
5. Masukan data ke salah satu klaster.

6. Hitung titik pusat *centroid* baru dengan persamaan $c_i = \sum_{j=1}^{n(s_i)} m_{ij} \ \epsilon \ s_i$ (5)
7. STOP.

Tabel 2. *Pseudocode* algoritma *k-means* dinamis dengan *mean based*

Input:

k: jumlah klaster, inisiasi klaster dinamis dinamis awal k = 2
mean: nilai *centroid* awal berdasarkan hasil metode *mean*
D: {d_1, d_2,…..,d_n} data set berisi n objek
changed: *boolean*

Output:

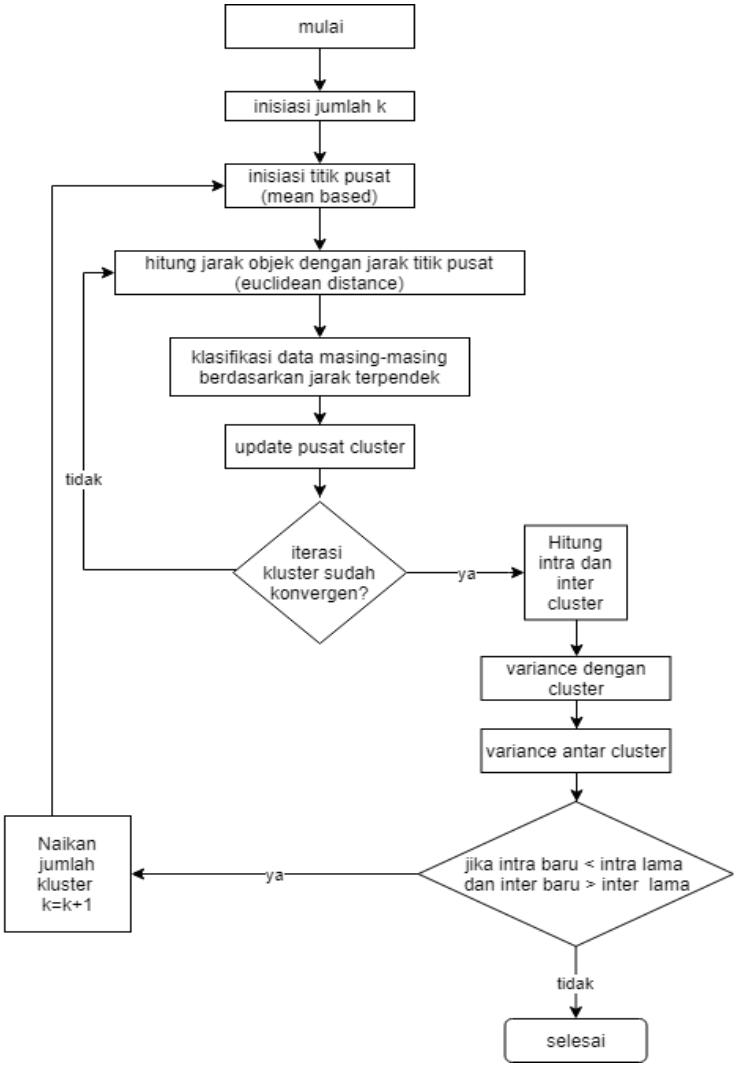
c: {c_1,c_2,….,c_k} nilai *centroid* klaster
L: {*l*(e) | e = 1,2,...,n} anggota klaster dari D

Method:

Inisiasi *centroid* awal menggunakan hasil nilai perhitungan *mean based* pada algoritma Tabel 1

1. Membuat partisi sejumlah 2 klaster dari segmentasi yang akan dibentuk.
2. Lakukan inisiasi titik pusat klaster hasil perhitungan *mean based*.
3. Hitung e*uclidean distance* dengan pusat klaster *mean* menggunakan persamaan 3 dan 4.
4. Perbarui anggota setiap klaster sesuai segmen terdekat. Hitung nilai rata-rata objek untuk setiap kluster.
5. Ulangi langkah 3-4 hingga data di segmentasi tidak berubah.
6. Jika fixed_no_of_clusters = true, pergi ke langkah 12.
7. Hitung jarak inter *cluster* menggunakan persamaan 1.
8. Hitung jarak intra *cluster* menggunakan persamaan 2.
9. Jika jarak intra *cluster* baru < jarak intra *cluster* lama dan jarak inter *cluster* baru > jarak inter *cluster* lama maka lanjut ke langkah 10. Jika tidak, maka pergi ke langkah 11.
10. k = k + 1, kembali ke langkah 1.
11. STOP.
12. Hasil pengelompokan data.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 9, Nomor 2, Juli 2020



Gambar 3. Algoritma klaster dinamis pada *k-means*

Algoritma klastering dinamis pada *k-means* memiliki kelebihan dalam melakukan pengecekan ulang kualitas klaster pada setiap iterasi, memungkinkan terjadinya perubahan jumlah klaster untuk memenuhi validitas kualitas klaster sehingga menghasilkan klaster yang optimal [2]. Diharapkan melalui penerapan optimasi *centroid* awal menggunakan metode *mean* menghasilkan hasil klaster menjadi lebih baik dan proses iterasi lebih optimal. Gambar 3 merupakan *flowcart* perancangan algoritma klaster dinamis menggunakan inisiasi *centroid* awal berbasis *mean*.

D. Teknik Analisis Data dan Evaluasi

Analisis dilakukan sejak penentuan data, proses *pra-processing*, dan implementasi perhitungan algoritma. Evaluasi klastering diuji tingkat validitasnya menggunakan beberapa metode yaitu *Cluster Variance* (V)*, Davies Bound Index* (DBI)*, Partition Coefficient* (PC)*,* dan *Sum Squared Error* (SSE)*.*

Untuk mengetahui evaluasi dari model yang digunakan, dilakukan dengan beberapa skenario percobaan untuk menentukan algoritma yang paling sesuai dan akurat dalam evaluasi kinerja akademik. Setelah semua hasil klaster terbentuk maka algoritma dibandingkan dan ditarik kesimpulan algoritma yang bekerja optimal, akurasi algoritma terbaik,

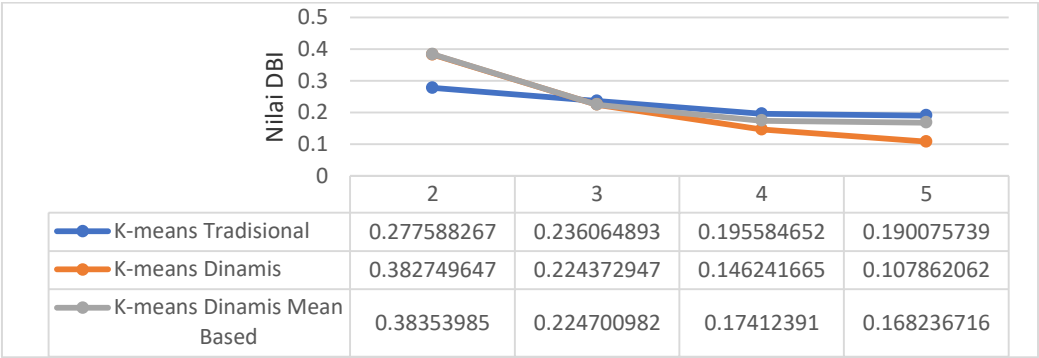
dan jumlah klaster yang paling optimal berdasarkan kriteria data uji.

HASIL DAN PEMBAHASAN

Implementasi algoritma *k-means* menggunakan basis data MySQL dan bahasa pemrograman PHP dengan perhitungan *k-means clustering* menggunakan data utama yaitu nilai UN, nilai rapor, dan perkembangan kinerja akademik melalui Indeks Prestasi Komulatif (IPK) semester 1 sampai semester 6. Data pendukung lain akan digunakan sebagai informasi tambahan dalam analisa klaster.

Untuk mengetahui perbandingan kinerja algoritma, skenario pengujian dilakukan sebanyak 10 kali secara bergantian. Uji coba *k-means* tradisional dilakukan dengan jumlah *cluster* yang beragam mulai 2 sampai 5 *cluster*. Sedangkan untuk algoritma *k-means* dinamis dan *k-means* dinamis dengan inisiasi *centroid* awal berbasis *mean* menggunakan hasil jumlah klaster ideal berdasarkan perhitungan *intra* dan *inter* klaster dari masing algoritma klaster dinamis.

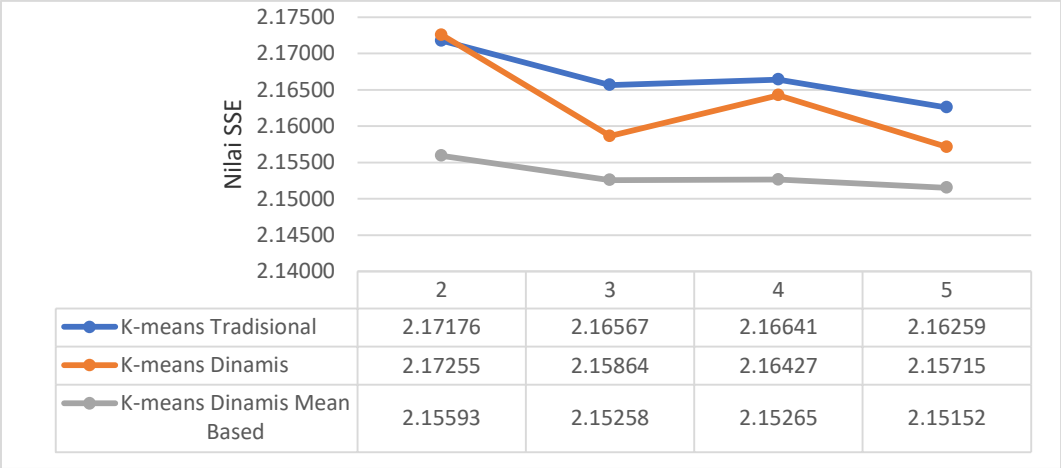
Penetapan titik pusat *cluster* awal (m_i) pada proses klastering *k-means* dinamis menggunakan bilangan *random*, disesuaikan dengan hasil jumlah hasil evaluasi *cluster* (kelompok). Sedangkan untuk metode *k-means* tradisional menggunakan nilai titik pusat klaster sama dengan nilai titik *centroid* awal pada *k-means* dinamis. Penilaian hasil evaluasi akhir ditentukan dengan perolehan nilai rata-rata dari 10 kali percobaan yang telah dilakukan sebelumnya, kemudian dibandingkan mulai 2 hingga 5 jumlah klaster yang terbentuk. Gambar 4 sampai Gambar 7 merupakan grafik evaluasi nilai SSE, PC, DBI, dan *variance* dari rata-rata 10 kali pengujian ketiga algoritma.



	2	3	4	5
K-means Tradisional	0.277588267	0.236064893	0.195584652	0.190075739
K-means Dinamis	0.382749647	0.224372947	0.146241665	0.107862062
K-means Dinamis Mean Based	0.38353985	0.224700982	0.17412391	0.168236716

Gambar 4. Perbandingan rata-rata nilai dbi ketiga algoritma

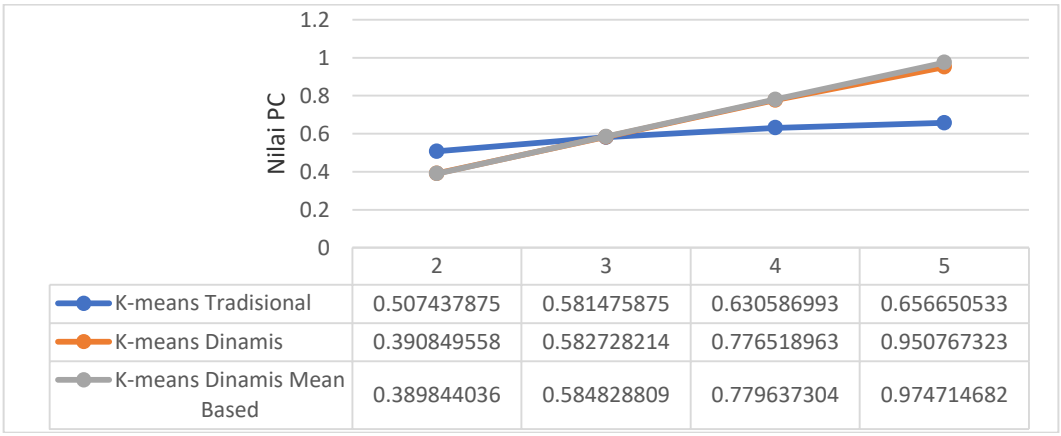
Secara keseluruhan hasil evaluasi DBI diperoleh nilai terbaik pada algoritma *k-means* dinamis dengan nilai DBI sebesar 0,107862. Penambahan jumlah klaster mempengaruhi hasil evaluasi DBI dimana penurunan nilai DBI pada algoritma klaster dinamis lebih baik.



	2	3	4	5
K-means Tradisional	2.17176	2.16567	2.16641	2.16259
K-means Dinamis	2.17255	2.15864	2.16427	2.15715
K-means Dinamis Mean Based	2.15593	2.15258	2.15265	2.15152

Gambar 5. Perbandingan rata-rata nilai sse ketiga algoritma

Pada pengujian rata-rata nilai SSE, algoritma *k-means* dinamis dengan inisiasi *centroid* awal berbasis *mean* memiliki performa yang lebih baik dibandingkan algoritma *k-means* tradisional dan *k-means* dinamis. Hal ini dapat dilihat dari perolehan nilai SSE yang selalu menunjukan nilai terkecil (mendekati 0) pada akhir iterasi dan jumlah klaster yang terbentuk selalu konsisten yaitu 5. Berdasarkan Gambar 5, nilai rata-rata SSE algoritma *k-means* tradisional dan *k-means* dinamis tidak stabil pada masing-masing jumlah klaster yang terbentuk. Secara keseluruhan nilai SSE terbaik (mendekati 0) diperoleh nilai 2,15152 pada algoritma *k-means* dinamis dengan inisiasi *centroid* berbasis *mean.*

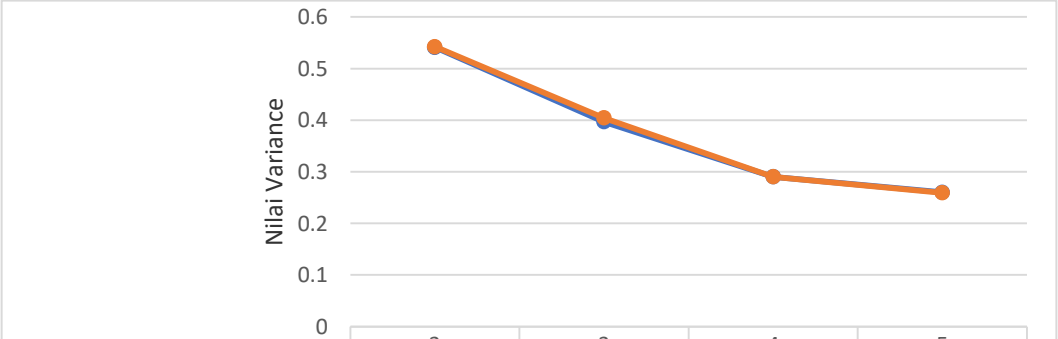


	2	3	4	5
K-means Tradisional	0.507437875	0.581475875	0.630586993	0.656650533
K-means Dinamis	0.390849558	0.582728214	0.776518963	0.950767323
K-means Dinamis Mean Based	0.389844036	0.584828809	0.779637304	0.974714682

Gambar 6. Perbandingan rata-rata nilai pc ketiga algoritma

Berdasarkan Gambar 6, seiring bertambahnya jumlah klaster maka nilai *purity* pada algoritma *k-means* dinamis dan *k-means* dinamis dengan inisiasi *centroid* mengalami peningkatan signifikan. Nilai *purity* terbesar (mendekati 1) diperoleh algoritma *k-means* dinamis dengan inisiasi *centroid* awal berbasis *mean* dengan nilai rata-rata tertinggi 0,974714 pada jumlah klaster 5 dan selisih 0,023947 lebih besar dari algoritma *k-means* dinamis. Apabila jumlah *cluster* semakin besar, maka nilai *purity k-means* dinamis dan *k-means* dinamis dengan inisiasi *centroid* akan semakin besar mendekati 1. Artinya hampir disetiap *cluster* selalu dihasilkan anggota kelompok yang selalu mirip dengan anggota yang lainnya.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 9, Nomor 2, Juli 2020



	2	3	4	5
K-means Dinamis	0.54128965	0.396981313	0.290245882	0.2601814
K-means Dinamis Mean Based	0.542407257	0.404423343	0.290210744	0.259281571

Gambar 7. Perbandingan rata-rata nilai *variance* klaster dinamis

Berdasarkan hasil evaluasi DBI, SSE, dan *variance*, terlihat jelas bahwa perubahan nilai masing-masing algoritma tergantung inisiasi *centroid* pada penentuan klaster awal yang digunakan. Algoritma *k-means* dinamis dengan inisiasi *centroid* memiliki kelebihan jumlah iterasi yang konstan pada setiap pengujian. Sehingga setiap dilakukan pengelompokan data, maka metode *k-means* dinamis dengan inisiasi *centroid* selalu menghasilkan klaster yang sama. Inisialisasi nilai titik pusat klaster awal yang dilakukan secara acak menyebabkan algoritma *k-means* tradisional dan *k-means* dinamis memperoleh nilai validasi dan jumlah klaster yang berbeda-beda disetiap pengujian, sehingga sangat sulit untuk memperoleh hasil *cluster* awal yang unik. Walaupun pengujian *k-means* dilakukan pengulangan sebanyak 10 kali, hasil pengujian tersebut belum tentu merepresentasikan suatu *cluster* yang baik. Akibatnya *centroid* yang diperoleh bukanlah *centroid* yang didominasi dengan pengelompokan kinerja akademik yang memiliki lebih banyak kemiripan. Proses iterasi setelah pembangkitan *cluster* awal akan menghasilkan *centroid* yang kurang tepat. Dengan demikian *cluster* yang akan terbentuk pada akhir bukanlah *cluster* yang memiliki nilai *purity* mendekati 1. Penentuan inisiasi klaster awal sangat mempengaruhi jumlah iterasi algoritma klastering dan estimasi waktu tidak konstan.

Hasil implementasi algoritma *k-means* dinamis dengan inisialisasi *centroid* menghasilkan jumlah klaster yang tetap pada setiap pengujian. Tabel 3 menjelaskan nilai titik pusat klaster awal metode *k-means* dengan inisiasi *centroid* awal berbasis *mean.*

Tabel 3. Nilai titik pusat inisialisasi k-means dinamis dengan *mean based*

Klaster	Var 1	Var 2	Var 3
1	3,545	28,837	81,841
2	3,526	50,242	81,751
3	3,574	61,202	82,467
4	3,629	72,150	84,114
5	3,575	84,797	83,950

Berdasarkan hasil pengujian, algoritma *k-means* dengan klaster dinamis menghasilkan kualitas klaster yang lebih baik dibandingkan dengan *k-means* tradisional [1]. Untuk mengetahui jumlah klaster yang paling baik pada algoritma *k-means* dinamis dengan inisialisasi klaster dilakukan pengujian terhadap tingkat validitas masing-masing jumlah klaster. Tabel 4 merupakan perbandingan masing-masing evaluasi pada masing-masing jumlah klaster. Kombinasi algoritma *k-means* klaster dinamis dengan inisialisasi klaster menetapkan dua klaster pada proses awal. Kemudian algoritma akan menghitung kembali klaster yang

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 9, Nomor 2, Juli 2020

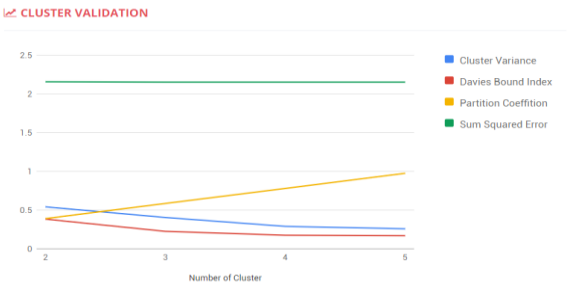
dihasilkan dan melakukan proses pengecekan terhadap nilai *variance* dari VW dan VB.

Tabel 4. Nilai validasi *k-means* klaster dinamis dengan *mean based*

No	Jumlah Klaster	VW	VB	DBI	PC	SSE	Iterasi
1	2	3.110633	5.734866	0.38354	0.389844	2.15593	14
2	3	2.436023	6.023448	0.224701	0.584829	2.152578	26
3	4	2.17997	7.51168	0.174124	0.779637	2.152648	30
4	5	1.935903	7.466411	0.168237	0.974715	2.15152	32

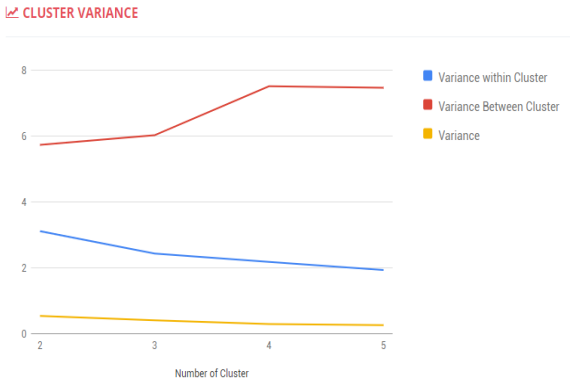
Berdasarkan Tabel 4, nilai VW jumlah 3 klaster lebih kecil dari jumlah klaster 2, sedangkan nilai VB jumlah klaster 3 lebih besar dari nilai VB jumlah klaster 2. Hal tersebut mengakibatkan algoritma *k-means* dinamis menambah satu klaster lagi menjadi klaster 4 sesuai hasil evaluasi klaster. Evaluasi algoritma *k-means* dinamis berhenti sampai pada jumlah klaster 5, dikarenakan nilai VW dan VB pada jumlah klaster 5 lebih kecil dari nilai VW dan VB pada jumlah 4 klaster seperti terlihat pada Gambar 8.

Jika dilihat dari nilai *Variance*, PC, dan SSE, algoritma klaster dinamis menggunakan inisiasi *centroid* berbasis *mean* memiliki jumlah klaster 5 yang lebih baik daripada jumah klaster 2, 3, dan 4. Hal tersebut dapat dilihat dari nilai SSE jumlah klaster 5 memperoleh nilai paling kecil dan untuk nilai PC jumlah klaster 5 memperoleh nilai paling mendekati 1. Gambar 9 merupakan hasil visual validitas masing-masing jumlah klaster berdasarkan nilai inisiasi titik pusat akhir yang terlihat pada Tabel 3.



Gambar 8. Nilai *cluster variance k-means* dinamis dengan *mean based*

Secara keseluruhan berdasarkan penerapan 4 metode evaluasi, maka dapat diketahui bahwa masing-masing metode memiliki hasil yang berbeda. Apabila dikaitkan terhadap domain pengaruhnya dalam penentuan algoritma terbaik, maka evaluasi menggunakan SSE menampilkan perbandingan hasil yang paling signifikan. Metode SSE dianggap mampu mengetahui evaluasi titik *centroid* yang telah dipilih dalam perhitungan klastering dan mengetahui kualitas persebaran masing-masing anggota klaster dengan melihat nilai *error* yang didapatkan. Selain itu hasil evaluasi nilai SSE selalu berada pada nilai normal (≥ 0) dan stabil, dibandingkan nilai PC yang dianggap semakin baik jika mendekati 1. Namun pada beberapa pengujian menghasilkan PC melebihi nilai 1.



Gambar 9. Nilai validitas jumlah klaster *k-means* dinamis dengan *mean based*

Apabila melakukan evaluasi jumlah klaster yang terbentuk, maka metode *cluster variance* dianggap paling cocok karena dapat mengelompokkan data observasi dalam jumlah besar dan variabel yang relatif banyak, sehingga data yang direduksi dengan kelompok akan mudah di analisis [8]. Selain itu dapat dipakai dalam skala data ordinal, interval, dan rasio [8]. Namun metode ini hanya dapat digunakan pada data yang bersifat *unsupervised* untuk

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 9, Nomor 2, Juli 2020

melihat hasil variasi penyebaran data yang ada pada sebuah klaster. Hasil selisih evaluasi klaster dengan *cluster variance* pada penelitian ini tidak terlalu jauh dengan jumlah klaster lainnya.

Analisis

Berdasarkan hasil klastering data kinerja akademik mahasiswa angkatan 2017 menggunakan algoritma *k-means* dinamis dengan inisialisasi klaster awal berbasis *mean*, didapatkan hasil 5 buah klaster ideal. Hasil jumlah klaster kemudian dapat dijadikan referensi untuk mengetahui persebaran kemampuan mahasiswa, untuk kemudian dikaitkan dengan jurusan yang ditempuh mahasiswa dalam bentuk prosentase.

Masing-masing anggota klaster memiliki nilai yang bervariasi dan sebaran data berbeda, anggota klaster 1 merupakan klaster dengan nilai titik pusat terendah, begitu seterusnya hingga klaster 5 dengan nilai titik pusat klaster tertinggi. Nilai tersebut dipengaruhi oleh hasil inisialisasi klaster awal pada Tabel 3, dimana inisialisasi menggunakan urutan rata-rata dari nilai terkecil ke paling besar pada masing-masing data. Gambar 10 merupakan hasil sebaran profil 765 data hasil klastering dinamis berbasis inisiasi *centroid* awal.

Gambar 10. Hasil profil klastering *k-means* dinamis dengan *mean based*

SIMPULAN

Algoritma klaster dinamis pada *k-means* dengan inisiasi *centroid* awal berbasis *mean* menghasilkan rata-rata kualitas klaster yang lebih baik dibanding algoritma lain dalam evaluasi *Sum Squared Error* (SSE), *Partition Coefficient* (PC) dan *Cluster Variance* (V). Keberhasilan algoritma dapat dilihat dari klaster yang dihasilkan berjumlah 5 klaster dengan nilai VW terkecil dengan nilai 1,935903 dan *variance* terkecil 0.259282 dibandingkan kedua algoritma lainnya yaitu *k-means* tradisional dan *k-means dinamis*.

Kelebihan algoritma *k-means* dinamis dengan inisiasi *centroid* adalah memiliki

jumlah interasi yang konstan pada setiap pengujian. Sehingga setiap kali pengelompokan data dilakukan maka metode *k-means* dinamis dengan inisiasi *centroid* selalu menghasilkan klaster yang sama. Sedangkan untuk algoritma *k-means* tradisional dan *k-means* klaster dinamis memperoleh nilai validasi dan jumlah klaster yang berbeda-beda disetiap pengujian dikarenakan inisialisasi nilai titik pusat klaster awal dilakukan secara acak sehingga sangat sulit untuk memperoleh hasil *cluster* awal yang unik dan jumlah iterasi yang kecil.

Jumlah klaster memberikan pengaruh terhadap nilai PC, SSE dan DBI. Jika jumlah klaster bertambah, maka nilai *purity* semakin mendekati nilai 1. Jika jumlah klaster berkurang, maka nilai SSE semakin membesar. Semakin banyak jumlah klaster dan data yang digunakan cenderung akan menghasilkan kualitas klaster atau pengelompokan yang semakin baik. Penambahan jumlah klaster mempengaruhi hasil evaluasi DBI ke arah lebih baik, walaupun algoritma lain tetap ada perbaikan hasil klastering, namun penurunan DBI pada algoritma *k-means* tradisional dan *k-means* dinamis dengan inisiasi *centroid* belum signifikan.

SARAN

Algoritma *k-means* dinamis dalam inisiasi titik pusat awal perlu dilakukan pengembangan atau pendekatan lain dalam menentukan dan memilih titik pusat *cluster* yang lebih baik untuk jenis data yang sama ataupun jenis data lainnya. Evaluasi kinerja akademik pada penelitian hanya digunakan sebagai data uji dan belum dilakukan analisis lebih lanjut sesuai kebutuhan dibidang akademik mendatang. Peneliti lain dapat menggunakan metode analisis *cluster* yang lebih baik dan dikembangkan dengan mengaplikasikan sesuai perkembangan algoritma terkini. Sehingga diharapkan mendapatkan titik pusat klaster dengan nilai *Davies Bound Index* (DBI*)* dan *Sum of Squared Error* (SSE) paling minimum.

REFERENSI

[1] Widiarina and R. Satria Wahono, “Algoritma Cluster Dinamik untuk Optimasi Cluster pada Algoritma K-Means dalam Pemetaan Nasabah Potensial,” *Journal of Intelligent Systems*, vol. 1, no. 1, pp. 33–36, 2015, [Online]. Available: http://journal.ilmukomputer.org.

[2] G. Akbari, Kerlooza, and Yusrila, “Peningkatan Hasil Cluster Menggunakan Algoritma Dynamic K-means dan K-means Binary Search Centroid,” *Jurnal Tata Kelola dan Kerangka Kerja Teknologi Informasi*, pp. 25–33, 2018.

[3] Narwati, “Pengelompokan Mahasiswa Menggunakan Algoritma K-Means,” *Jurnal Dinamika Informatika*, vol. 2, no. 2, 2010.

[4] G. I. Marthasari, “Implementasi Teknik Data Mining untuk Evaluasi Kinerja Mahasiswa Berdasarkan Data Akademik,” *Fountain of Informatics Journal*, vol. 2, no. 2, pp. 20–27, Nov. 2017, doi: 10.21111/fij.v2i2.1216.

[5] Sartikha, Maria, F. Winda Sari, and N. Jannah, “Analisis Profil Mahasiswa Politeknik Negeri Batam dengan Teknik Data Mining Asosiasi dan Clustering,” *Jurnal Integrasi*, vol. 8, no. 1, pp. 16–21, 2016.

[6] K. A. Seputra, I. Made Sudarma, and L. Jasa, “The Optimization of the Dynamic K-Means Clustering Algorithm with the Cluster Initialization in Grouping Travelers Perception to the Beach Tourist Destinations in Bali, Indonesia,” *International Journal of Research in IT*, vol. 07, no. 04, pp. 1–7, 2017, [Online]. Available: http://indusedu.org.

[7] S. B. M. Ahamed and K. S. Hareesha, “Dynamic Clustering of Data with Modified K-Means Algorithm,” *International Conference on Information and Computer*

Networks (ICICN 2012), vol. 27, pp. 221–225, 2012.

[8] I. T. Raharto, *Analisis Cluster: Tugas Mata Kuliah Konsep dan Teknik Perencanaan*. Makasar: Program Pascasarjana Universitas Hasanuddin, 2008.