Techno.COM, Vol. 19, No. 4, November 2020: 341-352

Pendekatan Initial Centroid Search Untuk Meningkatkan Efisiensi Iterasi Klustering K-Means

Initial Centroid Search Approach to Improve The Efficiency of K-Means Clustering Iterations

Muhammad Zulfahmi Nasution[1], Muhammad Siddik Hasibuan[2]
[1]Program Studi Sistem Komputer, Fakultas Sains dan Teknologi, Universitas Pembangunan Pancabudi, Medan Indonesia
[2]Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sumatera Utara, Medan Indonesia
E-mail: [1]fhm.nasution@dosen.pancabudi.ac.id, [2]muhammadsiddik@uinsu.ac.id

Abstrak

Pengelompokan K-Means bertujuan untuk mengumpulkan satu set titik pusat cluster yang optimal melalui iterasi yang berurutan. Fakta bahwa semakin optimal posisi dari titik pusat awal maka semakin sedikit jumlah iterasi dari algoritma pengelompokkan K-Means untuk konvergen. Oleh karena itu, Salah satu cara untuk menemukan set initial centroid adalah melalui metode iteratif guna mencari sejumlah initial centroid yang lebih baik untuk proses pengelompokan K-Means. Langkah awal yang kami lakukan adalah mengambil sampel data dari set data dan menjalankan algoritma K-Means sebagai proses awal untuk inisialisasi centroid cluster. Kemudian kami mengulang proses iterasi dengan sejumlah initial centroid yang telah diinisialisasikan sebelumnya dan mengukur hasil pengelompokkan melalui *sum-of-square-error* guna menentukan kebaikan keanggotaan cluster. Centroid akhir yang memberikan jarak terendah yang akan kami teruskan ke proses pengelompokan K-means secara lengkap. Harapan kami adalah pendekatan ini akan mengarah pada set initial centroid yang lebih baik sebagai proses pengelompokan K-Means sehingga mampu meningkatkan kinerja Algoritma K-Means karena hasil konvergensi Algoritma K-Means akan berbanding lurus dengan pemilihan *initial centroid.*

Kata kunci: Titik Pusat Cluster, Pengelompokkan, Sum of Square Error, Konvergen, K-Means

Abstract

K-means clustering aims to gather a set of optimal initial centroid through successive iterations. The fact that more optimal the point of the initial cluster centers, the less iteration of the K-Means clustering algorithm will be needed for convergence. Therefore, one way to find a better initial set of centroid is through an iterative approach to searching for a better set of initial centroid for K-Means clustering. The first step we will take is to take data samples from the data set and run the short runs of the K-Means clustering algorithm on it (not for convergence) but as the initial process of centroids initialization. Then we will repeat the short runs as an iteration process with a number of initial centroid cluster being randomly initialized before and measuring within-cluster by sum-of-squares-error to determine the goodness of cluster membership. The final centroids that provides the lowest inertia will continue to complete the k-means grouping process. Our hope is that this approach will lead to a better initial centroid set for the k-means clustering to improve the performance of the K-Means Algorithm because the convergence results of the K-Means Algorithm will be directly proportional to the selection of initial centroids.

Keywords: Initial Centroid, Clustering, Sum of Square Error, Convergence, K-Means

1. PENDAHULUAN

Semakin banyaknya volume data yang terus tumbuh dalam kompleksitas dan keragaman yang didukung oleh trend perangkat keras dan perangkat lunak aplikasi yang semakin canggih menjadikan semakin sulitnya untuk mengekstraksi data mentah menjadi informasi yang bermanfaat. Saat ukuran data semakin melampaui kemampuan manusia untuk menemukan pola data dengan berbagai karakteristik dan label dari sejumlah besar data yang bersifat massal, maka saat itulah teknologi komputasi dibutuhkan untuk melakukan proses penggalian informasi yang bermanfaat. Salah satu pendekatan komputasi yang diperlukan adalah Data Mining.

Data Mining merupakan teknik penambangan data massal dengan cara mendapatkan pola dari data mentah menjadi informasi. Salah satu kegiatan yang paling umum dalam data mining adalah mengelompokkan data dari satu set data kedalam kategori atau kelompok tertentu. Dalam teknik pengelompokan, objek dikelompokkan atas dasar subjektif kemiripan objek.

Kemiripan antar objek dalam suatu kelompok memiliki nilai yang lebih tinggi daripada kemiripan antara objek dalam suatu kelompok yang berbeda. Teknik pengelompokan dalam data mining disebut Clustering. Clustering adalah teknik eksplorasi data dengan cara mengelompokkan objek-objek dengan karakteristik atau ciri yang serupa sehingga memudahkan pemrosesan pengelompokkan data mentah menjadi informasi yang lebih bermanfaat. Clustering bersifat Unsupervised Learning [1] Berbeda dengan Supervised Learning yang memiliki kelas atau label (actual output), Pada unsupervised learning hanya ada data masukan (input) dan parameter tanpa adanya kelas atau label (actual output). Metode Clustering terbagi atas empat kategori yaitu Partition Clustering Method, Hierarchical Clustering Method, Density-based Clustering Method dan Grid-based Clustering Method [2]. Namun, Pendekatan yang sering digunakan terdiri atas dua yakni Hierarchical Clustering dan Non-Hierarchical Clustering. Salah satu yang termasuk kedalam pendekatan Non-Hierarchical Clustering adalah Partition Clustering [3]. Metode Partition Clustering memiliki dua algoritma yaitu K-Medoids dan K-Means [4]. Algoritma K-Means termasuk kedalam metode Partition Clustering yang mampu mengelompokkan data dengan cara membagi partisi data ke dalam sejumlah kategori atau kelompok tertentu [5].

Algoritma K-Means untuk pertama kali diperkenalkan oleh MacQueen pada tahun 1967. Data yang memiliki selisih jarak terdekat dari titik pusat cluster akan dijadikan sebagai titik pusat cluster baru dengan cara menghitung rata-rata nilai dari data (selisih jarak terdekat) yang termasuk kedalam cluster tertentu. Algoritma K-Means mengulangi tahapan menghitung selisih jarak setiap data ke masing-masing titik pusat cluster baru sampai tidak ada data yang berpindah cluster (Convergence) atau sampai batas maksimal iterasi yang ditentukan [6].

Algoritma K-means konvensional [5] berupaya meminimalkan selisih jarak (nearest distance) antar data dan memperbarui titik pusat awal cluster pada setiap iterasi [7]. Hal ini yang menyebabkan data yang selalu berpindah cluster sampai batas maksimal iterasi yang ditentukan [8] Pada algoritma K-means konvensional, Penentuan titik pusat awal cluster masih dilakukan secara acak (random).

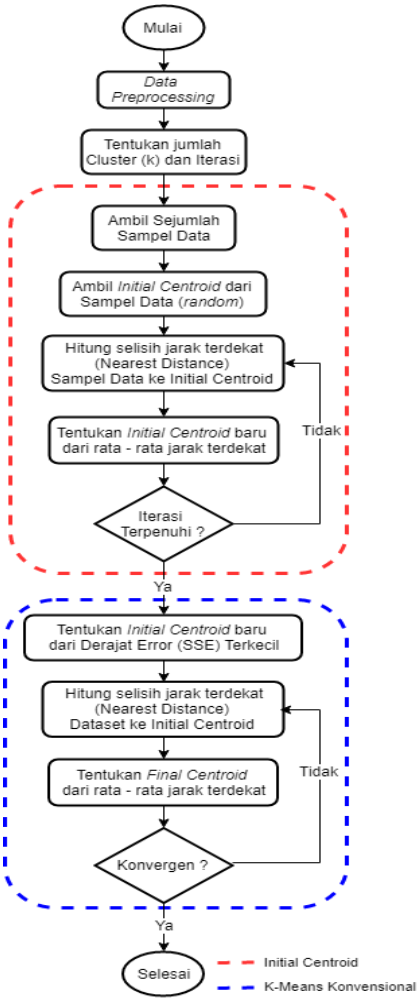
Konfigurasi titik pusat awal cluster yang berbeda dapat menyebabkan nilai centroid akhir yang berbeda pula dan hasil clustering menjadi tidak stabil serta tidak akurat [2] karena itu, Pemilihan nilai titik pusat awal cluster atau disebut initial centroid[9] sangatlah penting.

Tujuan dari dilakukannya penelitian ini adalah untuk memberikan solusi terhadap permasalahan dalam proses penentuan dan pemilihan nilai initial centroid yang masih dilakukan secara acak pada Algoritma K-means konvensional[10][11]. Dalam penelitian ini, penentuan dan pemilihan nilai initial centroid dilakukan menggunakan metode iteratif. Langkah awal yang dilakukan adalah mengambil sampel data dari set data dan menjalankan tahapan Algoritma K-means sebagai proses awal untuk inisialisasi centroid. Kemudian mengulang proses iterasi dengan sejumlah initial centroid yang telah diinisialisasikan sebelumnya.

Proses mengulang iterasi dilakukan untuk memperoleh initial centroid yang memberikan jarak terdekat dan derajat error terkecil yang akan diteruskan kedalam tahapan algoritma K-Means secara lengkap[12]. Sehingga item data mengarah kedalam cluster yang tepat dan meminimalkan data yang berpindah cluster. Metode yang diusulkan yakni Initial Centroid Search diharapkan mampu meningkatkan kinerja Algoritma K-Means karena hasil konvergensi Algoritma K-Means akan berbanding lurus dengan pemilihan initial centroid[13].

2. METODE PENELITIAN

Metode clustering yang diusulkan memiliki perbedaan tahapan dari Algoritma K-means konvensional. Dalam penelitian ini, penentuan dan pemilihan nilai initial centroid dilakukan menggunakan metode iteratif. Langkah awal yang dilakukan adalah mengambil sampel data dari set data dan menjalankan tahapan Algoritma K-means sebagai proses awal untuk inisialisasi centroid. Untuk lebih rinci dan jelas dalam mendeskripsikan proses dalam penelitian ini, M¬aka berikut adalah tahapan demi tahapan pada metode usulan yaitu:



Gambar 1. Tahapan Metode Usulan

Berdasarkan Gambar 2, Tahapan netode usulan dapat diuraikan sebagai berikut :

1. Melakukan 2 proses data preprocessing yaitu proses cleaning untuk membuang duplikasi data dan proses normalization agar interval atau rentang data menjadi lebih proporsional.

2. Menentukan sejumlah cluster (k) dan jumlah iterasi.
3. Mengambil sejumlah sampel dari set data.
4. Menentukan initial centroid secara acak dari sampel.
5. Menghitung selisih jarak terdekat dari sampel ke initial centroid menggunakan perhitungan jarak (Euclidean Distance).
6. Menentukan cluster sampel dengan jarak terdekat ke initial centroid
7. Menentukan initial centroid baru berdasarkan nilai rata-rata cluster sampel dengan jarak terdekat ke initial centroid
8. Mengulangi tahapan nomor 5, Jika iterasi belum terpenuhi. Sebaliknya, iterasi terpenuhi maka lanjut ke tahapan berikutnya.
9. Menentukan initial centroid baru dengan derajat error (SSE) terkecil.

Tahapan selanjutnya melakukan *clustering* K-Means konvensional menggunakan *initial centroid* yang telah diperoleh sebelumnya dari metode usulan sampai tidak ada data yang berpindah cluster (*Convergen*).

3. HASIL DAN PEMBAHASAN

A. Data yang Digunakan

Untuk menguji kinerja metode yang diusulkan maka digunakanlah dua set data dari Badan Pusat Statistik Sumatera Utara (https://sumut.bps.go.id) yang terdiri yaitu Persentase penduduk yang mempunyai Keluhan Kesehatan selama sebulan terakhir menurut Kabupaten/Kota di Provinsi Sumatera Utara tahun 2014 – 2019 dan Persentase Rumah Tangga yang memiliki akses terhadap Sanitasi Layak menurut Kabupaten/Kota di Provinsi Sumatera Utara tahun 2014 – 2019. Berikut adalah informasi rinci dataset persentase penduduk yang mempunyai Keluhan Kesehatan pada tabel berikut :

Tabel 1. Rincian Dataset Persentase Penduduk yang Mempunyai Keluhan Kesehatan Sebulan Terakhir menurut Kabupaten/Kota Di Provinsi Sumatera Utara Periode 2014 – 2019

Atribut	**Nilai**
Kabupaten/Kota	Sumatera Utara,Nias,Madina,Tapanuli Selatan,Tapanuli Tengah,Tapanuli Utara,Toba Samosir,LabuhanBatu,Asahan,Simalungun,Dairi,Karo,DeliSerdang,Langkat,NiasSelatan,Humbang Hasundutan,Pakpak Bharat,Samosir,Serdang Bedagai,BatuBara,Padang Lawas Utara,Padang Lawas,LabuhanbatuSelatan,LabuanBatuUtara, NiasUtara,NiasBarat,Sibolga,Tanjungbalai, Pematangsiantar,TebingTinggi,Medan,Binjai, Padangsidimpuan,Gunungsitoli
Tahun 2014	[10.86 – 37.52]
Tahun 2015	[13.98 – 30.18]
Tahun 2016	[12.07 -33.35]
Tahun 2017	[12.30 -32.20]
Tahun 2018	[14.16 -35.97]
Tahun 2019	[13.96 -39.54]

Berdasarkan Tabel 1, dapat dilihat rincian dataset persentase penduduk yang mempunyai keluhan kesehatan sebulan terakhir menurut Kabupaten/Kota di Provinsi Sumatera Utara selama periode Tahun 2014 sampai 2019.

Berikut adalah informasi rincian dataset persentase rumah tangga yang memiliki akses terhadap sanitasi yang layak pada tabel berikut :

Tabel 2. Rincian Dataset Persentase Rumah Tangga yang memiliki Akses Sanitasi yang Layak menurut Kabupaten/Kota Di Provinsi Sumatera Utara
Periode 2014 – 2019

Atribut	**Nilai**
Kabupaten/Kota	Sumatera Utara,Nias,Madina,Tapanuli Selatan,Tapanuli Tengah,Tapanuli Utara,Toba Samosir,LabuhanBatu,Asahan,Simalungun,Dairi,Karo,DeliSerdang,Langkat,NiasSelatan,Humbang Hasundutan,Pakpak Bharat,Samosir,Serdang Bedagai,BatuBara,Padang Lawas Utara,Padang Lawas,LabuhanbatuSelatan,LabuanBatuUtara, NiasUtara,NiasBarat,Sibolga,Tanjungbalai, Pematangsiantar,TebingTinggi,Medan,Binjai, Padangsidimpuan,Gunungsitoli
Tahun 2014	[5.59 – 93.91]
Tahun 2015	[10.95 – 93.92]
Tahun 2016	[7.9 -94.42]
Tahun 2017	[3.72 -95.61]
Tahun 2018	[7.4 -95.38]
Tahun 2019	[10.55 -95.62]

Berdasarkan Tabel 2, dapat dilihat rincian dataset persentase rumah tangga yang memiliki akses sanitasi yang layak menurut Kabupaten/Kota Di Provinsi Sumatera Utara selama periode tahun 2014 sampai 2019.

A. Hasil Pemilihan Initial Centroid pada Metode Initial Centroid Search

1. Hasil pemilihan centroid awal terhadap dataset kelihan kesehatan

Berikut adalah hasil pemilihan i*nitial centroid* dari dataset Persentase Penduduk dengan Keluhan Kesehatan menggunakan metode usulan yakni *Initial Centroid Search* sebanyak 4 cluster pada gambar 2

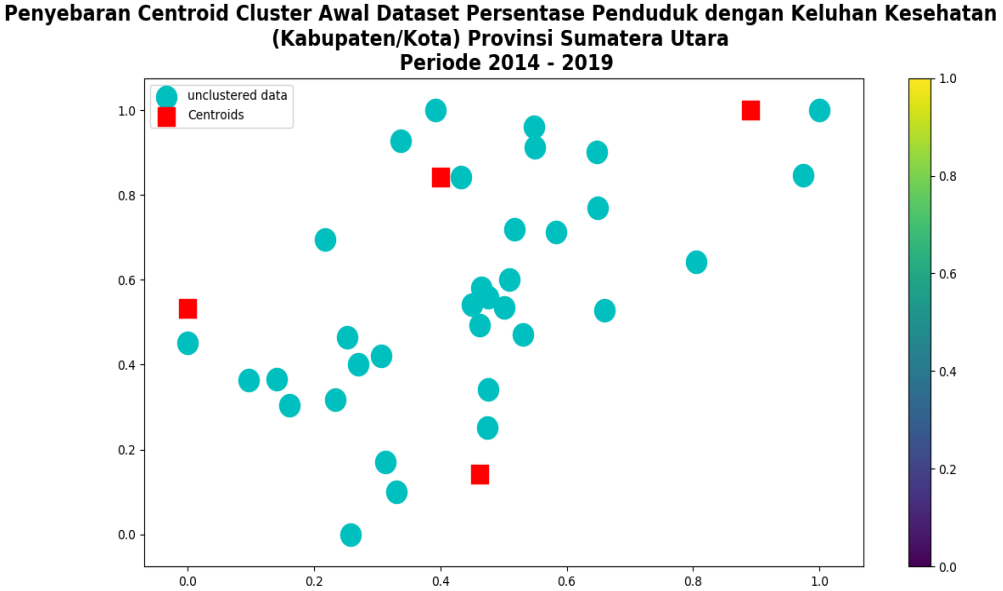
```
-------->>>>>> Proses Clustering dengan K = 4 .....

--Nilai Centroid Awal (Dataset Keluhan Kesehatan)--
       Cluster 1  Cluster 2  Cluster 3  Cluster 4
2014   0.400000   0.841975   0.385808   0.666823
2015   0.281015   0.376880   0.462491   0.141035
2016   0.270443   0.432858   0.492593   0.366667
2017   0.890417   1.000000   0.465989   0.643862
2018   0.905004   0.658327   0.000000   0.532663
2019   0.331658   0.472362   0.408987   0.302155
```

Gambar 2. Perolehan Nilai *Centroid* Awal Dataset Keluhan Kesehatan

Berdasarkan Gambar 2, merupakan hasil perolehan nilai *centroid* awal yang dihasilkan menggunakan metode *clustering Initial Centroid Search* terhadap dataset Persentase Penduduk dengan Keluhan Kesehatan. Nilai centroid awal periode Tahun 2014 secara berurutan pada cluster 1 senilai 0.40, cluster 2 senilai 0.84, cluster 3 senilai 0.38 dan cluster 4 senilai 0.66. Hal yang sama juga berlaku untuk periode tahun berikutnya.

Berikutnya adalah hasil penyebaran titik *centroid cluster* awal yang dihasilkan menggunakan metode *clustering Initial Centroid Search* terhadap dataset Persentase Penduduk dengan Keluhan Kesehatan berdasarkan jumlah cluster (K) sebanyak 4 terlihat pada gambar 3



Gambar 3. Sebaran *Initial Centroid* Persentase Penduduk dengan Keluhan Kesehatan

Dari gambar 3, menunjukkan hasil tampilan secara visual tentang sebaran *centroid cluster* awal metode *clustering Initial Centroid Search* saat nilai K=4 pada dataset Persentase Penduduk dengan Keluhan Kesehatan. Dapat dilihat simbol persegi berwarna merah menunjukkan titik *centroid cluster*. Sedangkan simbol lingkaran berwarna *cyan* menunjukkan titik nilai yang belum dilakukan *clustering* dari masing-masing atribut pada dataset Persentase Penduduk dengan Keluhan Kesehatan.

2. *Hasil pemilihan Centroid awal terhadap dataset akses sanitasi yang layak*

Berikut adalah hasil pemilihan nilai *centroid* awal masing-masing atribut dari dataset Persentase Rumah Tangga yang memiliki Akses Sanitasi yang layak menggunakan metode *clustering Initial Centroid Search* dimana jumlah cluster (K) sebanyak 4 cluster pada gambar 4

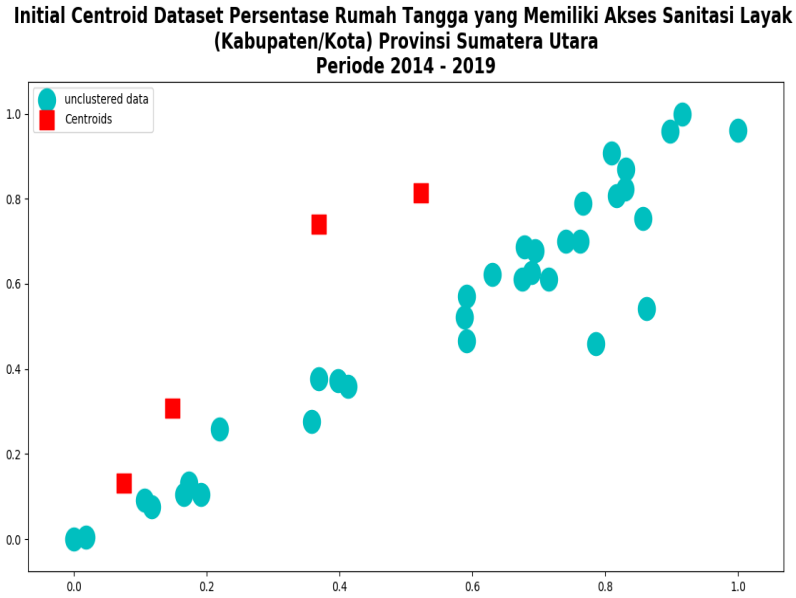
```
--------->>>>>>> Proses Clustering dengan K = 4 .....

--Nilai Centroid Awal (Dataset Akses Sanitasi Layak)--
      Cluster 1  Cluster 2  Cluster 3  Cluster 4
2014   0.368207   0.740489   0.116621   0.173120
2015   0.377124   0.700012   0.522037   0.814996
2016   0.148547   0.285232   0.467038   0.771880
2017   0.148102   0.307343   0.406019   0.844363
2018   0.084754   0.313859   0.075328   0.132337
2019   0.492256   0.753814   0.000000   0.226306
```

Gambar 4. Perolehan Nilai *Centroid* Awal Dataset Akses Sanitasi yang Layak

Berdasarkan Gambar 4, merupakan hasil perolehan nilai *centroid* awal yang dihasilkan menggunakan metode *clustering Initial Centroid Search* terhadap dataset Persentase Rumah Tangga dengan Akses Sanitasi yang layak. Nilai centroid awal periode Tahun 2014 secara

berurutan pada cluster 1 senilai 0.36, cluster 2 senilai 0.74, cluster 3 senilai 0.11 dan cluster 4 senilai 0.17. Hal yang sama juga berlaku untuk periode tahun berikutnya yang terlihat seperti gambar 5



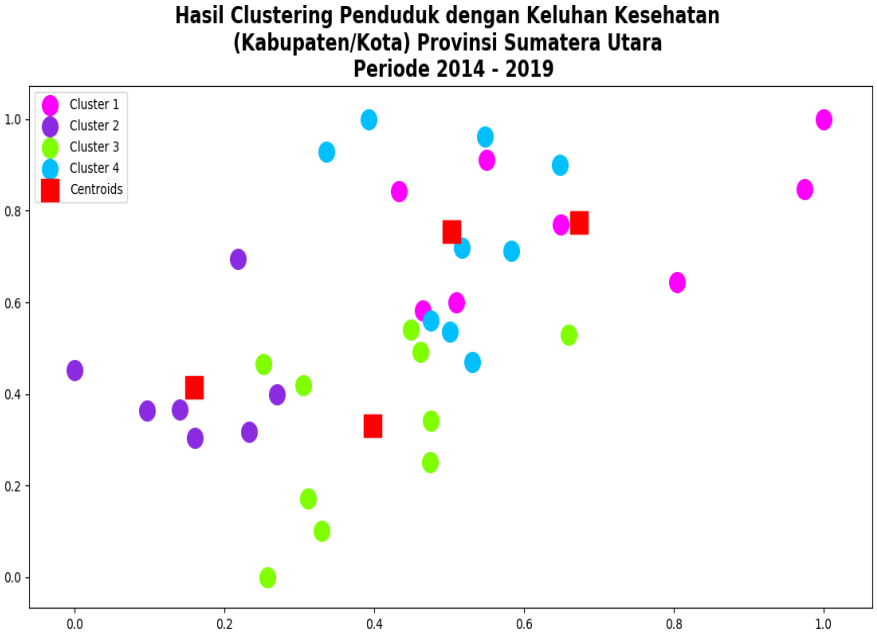
Gambar 5. Penyebaran *Centroid Cluster* Awal Dataset Akses Sanitasi yang Layak

Terlihat pada gambar 5 merupakan bentuk sebaran nilai *Initial Centroid* pada dataset Persentase Rumah Tangga yang memiliki Akses Sanitasi yang layak menggunakan metode *Initial Centroid Search.*

B. Hasil Konvergensi Clustering Clustering terhadap Dataset Keluhan Kesehatan

1. Hasil Konvergensi Clustering terhadap Dataset Keluhan Kesehatan

Berikut adalah hasil dari konvergensi *clustering* dataset Persentase Penduduk dengan Keluhan Kesehatan menggunakan metode *Initial Centroid Search* sebanyak 4 cluster pada gambar 6



Gambar 6. Tampilan Clustering Metode *Initial Centroid Search* terhadap dataset Persentase Penduduk dengan Keluhan Kesehatan

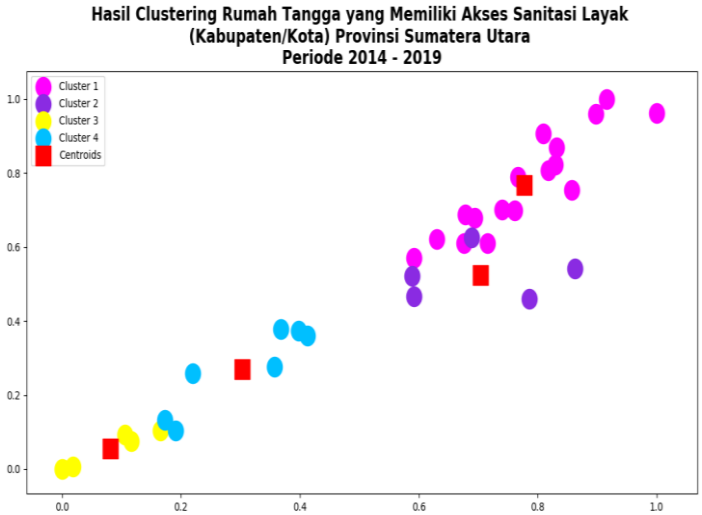
Terlihat pada gambar 4 merupakan bentuk konvergensi clustering dataset Persentase Penduduk dengan Keluhan Kesehatan yang dihasilkan oleh metode *Initial Centroid Search* sebanyak 4 cluster. Berikut adalah hasil *clustering* secara rinci:

Tabel 3. Hasil Clustering Persentase Penduduk yang Mempunyai Keluhan Kesehatan Kabupaten/Kota Di Provinsi Sumatera Utara Periode 2014 – 2019
(Metode *Initial Centroid Search)*

Kelompok	Nilai
Cluster 1	Sumatera Utara, Toba Samosir, Simalungun,Dairi,Karo,DeliSerdang,Langkat,AsahanHumbang Hasundutan, Samosir, Serdang Bedagai, BatuBara, Tanjungbalai, Pematangsiantar, TebingTinggi, Medan, Binjai
Cluster 2	Tapanuli Selatan, Tapanuli Tengah, Padang Lawas, Padang Lawas Utara, Sibolga, Padangsidimpuan, Gunungsitoli
Cluster 3	Nias, Madina, Nias Selatan, Nias Utara,Nias Barat
Cluster 4	LabuhanBatu, Pakpak Bharat, Labuhanbatu Selatan, Labuanbatu Utara, Tapanuli Utara

2. Dataset Akses Sanitasi yang Layak

Berikut adalah hasil dari konvergensi *clustering* dataset Persentase Rumah Tangga yang memiliki Akses Sanitasi yang layak menggunakan metode *Initial Centroid Search* sebanyak 4 cluster pada gambar 7



Gambar 7. Konvergensi Clustering pada Persentase Rumah Tangga yang memiliki Akses Sanitasi yang layak (Metode Initial Centroid Search)

Terlihat pada gambar 5 merupakan bentuk konvergensi *clustering* dataset Persentase Rumah Tangga yang memiliki Akses Sanitasi layak yang dihasilkan oleh metode *Initial Centroid Search* sebanyak 4 cluster. Berikut adalah hasil *clustering* secara rinci:

Tabel 4. Hasil Clustering Persentase Rumah Tangga yang memiliki Akses Sanitasi yang Layak menurut Kabupaten/Kota Di Provinsi Sumatera Utara Periode 2014 – 2019 (Metode *Initial Centroid Search)*

Kelompok	Nilai
Cluster 1	Nias, Madina, Nias Selatan, Nias Utara,Nias Barat
Cluster 2	Tapanuli Selatan, Tapanuli Tengah, Padang Lawas, Padang Lawas Utara, Sibolga, Padangsidimpuan, Gunungsitoli
Cluster 3	LabuhanBatu, Pakpak Bharat, Labuhanbatu Selatan,

	Labuanbatu Utara, Tapanuli Utara
Cluster 4	Sumatera Utara, Toba Samosir, Simalungun,Dairi,Karo,DeliSerdang,Langkat, Asahan, Humbang Hasundutan, Samosir, Serdang Bedagai, BatuBara, Tanjungbalai, Pematangsiantar, TebingTinggi, Medan, Binjai

C. Hasil Pengujian Metode Clustering Initial Centroid Search dan K-Means Konvensional

Pengujian dilakukan menggunakan salah satu dataset yaitu Persentase Penduduk dengan Keluhan Kesehatan selama sebulan terakhir menurut Kabupaten/Kota di Provinsi Sumatera Utara dalam periode 2014 – 2019.

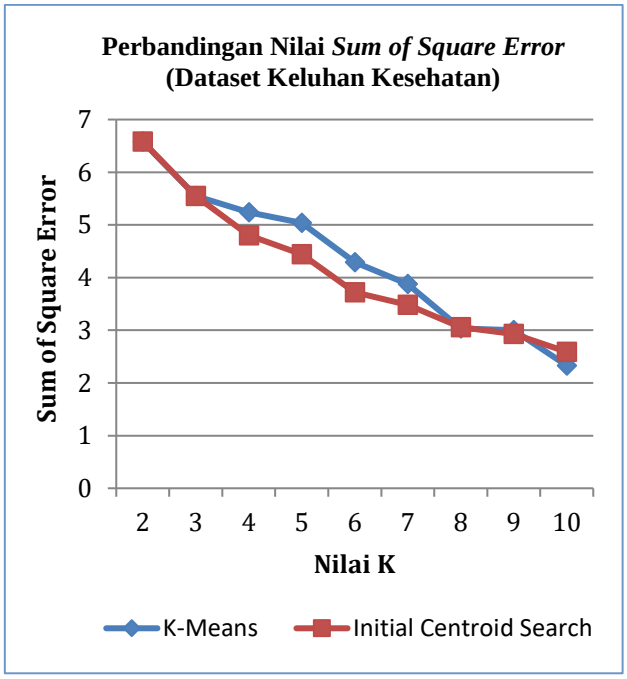
Berikut adalah tabel hasil pengujian metode *Initial Centroid Search* dan K-Means Konvensional menurut pengukuran *Sum of Square Error* (*SSE*), yaitu:

Tabel 5. Hasil Pengujian Metode Clustering *Initial Centroid Search* dengan K-Means Konvensional (Dataset Keluhan Kesehatan)

K	***Sum of Square Error***		**Metode Pilihan**
	K-Means(1)	**Metode Usulan (2)**	
2	6.58	6.58	(1)&(2)
3	5.55	5.55	(1)&(2)
4	5.24	4.80	(2)
5	5.04	4.44	(2)
6	4.29	3.72	(2)
7	3.88	3.48	(2)
8	3.03	3.06	(1)
9	3.00	2.93	(2)
10	2.33	2.59	(1)

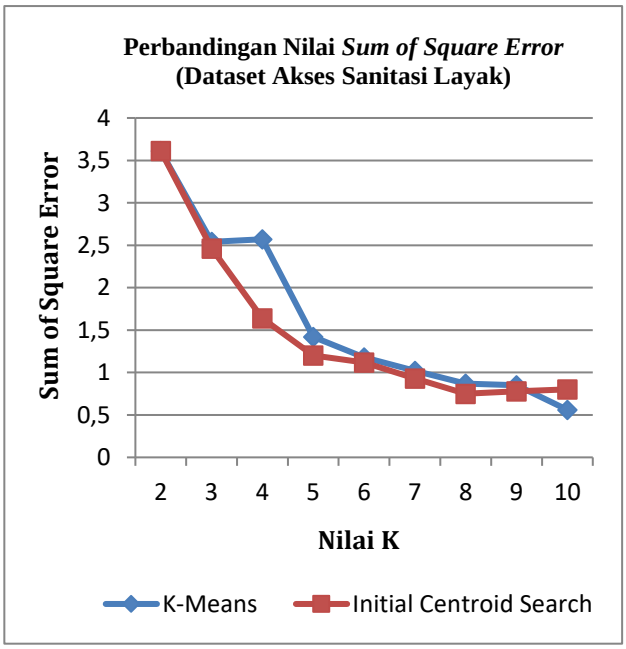
Berdasarkan informasi pada tabel 5, metode usulan telah mampu memberikan nilai derajat *error* (*Sum of Square Error*) yang minimal dari metode K-Means konvensional. Metode *Initial Centroid Search* mayoritas memberikan nilai error terkecil, terbukti saat jumlah cluster berada pada nilai K=4, K=5, K=6, K=7 dan K=9. Hal ini tidak lepas dari proses pemilihan *initial centroid* yang telah dihasilkan dari metode *Initial Centroid Search* seperti pada gambar 2.

Berikut adalah hasil perbandingan kinerja metode *Initial Centroid Search* dan K-Means konvensional dalam bentuk grafik berdasarkan pengukuran *Sum of Square Error* (*SSE*) menggunakan kedua dataset pada gambar 8 dan 9.



Gambar 8. Grafik Perbandingan Error Metode *Initial Centroid Search* dengan K-Means Konvensional (Dataset Persentase Keluhan Kesehatan)

Dari gambar 8, menunjukkan hasil perbandingan kinerja metode *clustering Initial Centroid Search* dengan K-Means konvensional yang diukur berdasarkan metode *Sum of Square Error*. Dapat dilihat bahwa nilai error terkecil secara signifikan terjadi pada metode *Initial Centroid Search* saat nilai K mulai berada pada K=4 sampai K=7 dengan selisih error dari kedua metode rata-rata sebesar 0.51 poin.



Gambar 9. Grafik Perbandingan Error Metode *Initial Centroid Search* dengan K-Means Konvensional (Dataset Persentase Akses Sanitasi Layak)

Dari gambar 10, menunjukkan hasil perbandingan kinerja metode *clustering Initial Centroid Search* dengan K-Means konvensional yang diukur berdasarkan metode *Sum of Square Error*. Dapat dilihat bahwa nilai error terkecil secara signifikan terjadi pada metode *Initial Centroid*

Search saat nilai K mulai berada pada K=4 dan K=5 dengan selisih error dari kedua metode rata-rata sebesar 0.57 poin.

4.KESIMPULAN

Proses mengulang iterasi dilakukan untuk memperoleh nilai *centroid* awal yang memberikan jarak terdekat sehingga derajat *error* menjadi kecil. Metode ini diusulkan agar hasil clustering mengarah kedalam keanggotaan cluster yang tepat dan meminimalkan data yang berpindah cluster. Metode *clustering Initial Centroid Search* terbukti menghasilkan nilai *error* terkecil saat nilai K mulai berada pada K=4 sampai K=9 seperti yang dapat dilihat pada gambar 8 dan 9. Hal ini membuktikan bahwa hasil konvergensi clustering Algoritma K-Means memang berbanding lurus dengan pemilihan nilai centroid awal (*initial centroid cluster)*.

UCAPAN TERIMA KASIH

Penghargaan tinggi sudah semestinya diberikan kepada segenap Lembaga Penelitian Universitas Pembangunan Pancabudi Medan (LPPM UNPAB), Rektor Universitas Pembangunan Pancabudi Medan, Fakultas Sains dan Teknologi, Program Studi Sistem Komputer dan LPPM Universitas Islam Negeri Sumatera Utara Fakultas Sains dan Teknologi Program Studi Ilmu Komputer, Atas dukungan terhadap karya penelitian ini dan fasilitas yang telah diberikan.

DAFTAR PUSTAKA

[1] J. Ortiz-Bejar, E. S. Tellez, M. Graff, J. Ortiz-Bejar, J. C. Jacobo, and A. Zamora-Mendez, "Performance analysis of k-means seeding algorithms," in *2019 IEEE International Autumn Meeting on Power, Electronics and Computing, ROPEC 2019*, 2019, doi: 10.1109/ROPEC48299.2019.9057044.

[2] C. Xiong, Z. Hua, K. Lv, and X. Li, "An improved K-means text clustering algorithm by optimizing initial cluster centers," in *Proceedings - 2016 7th International Conference on Cloud Computing and Big Data, CCBD 2016*, 2017, doi: 10.1109/CCBD.2016.059.

[3] Y. Chen, P. Hu, and W. Wang, "Improved K-Means Algorithm and its Implementation Based on Mean Shift," in *Proceedings - 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2018*, 2019, doi: 10.1109/CISP-BMEI.2018.8633100.

[4] V. Divya and K. N. Devi, "An Efficient Approach to Determine Number of Clusters Using Principal Component Analysis," in *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018*, 2018, doi: 10.1109/ICCTCT.2018.8551182.

[5] A. Ilham, D. Ibrahim, L. Assaffat, and A. Solichan, "Tackling Initial Centroid of K-Means with Distance Part (DP-KMeans)," in *Proceeding - 2018 International Symposium on Advanced Intelligent Informatics: Revolutionize Intelligent Informatics Spectrum for Humanity, SAIN 2018*, 2019, doi: 10.1109/SAIN.2018.8673364.

[6] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?," *Pattern Recognit.*, 2019, doi: 10.1016/j.patcog.2019.04.014.

[7] D. Tanir and F. Nuriyeva, "On selecting the initial cluster centers in the K-means algorithm," in *11th IEEE International Conference on Application of Information and Communication Technologies, AICT 2017 - Proceedings*, 2019, doi: 10.1109/ICAICT.2017.8687081.

[8] J. James Manoharan and S. Hari Ganesh, "Initialization of optimized K-means centroids using divide-and-conquer method," *ARPN J. Eng. Appl. Sci.*, 2016.

[9] C. M. Poteras, M. C. Mihaescu, and M. Mocanu, "An optimized version of the K-Means clustering algorithm," in *2014 Federated Conference on Computer Science and Information Systems, FedCSIS 2014*, 2014, doi: 10.15439/2014F258.

[10] G. Shi, B. Gao, and L. Zhang, "The optimized K-means algorithms for improving randomly-initialed midpoints," in *Proceedings of 2013 2nd International Conference on Measurement, Information and Control, ICMIC 2013*, 2013, doi: 10.1109/MIC.2013.6758177.

[11] M. Goyal and S. Kumar, "Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability," *J. Inst. Eng. Ser. B*, 2014, doi: 10.1007/s40031-014-0106-z.

[12] H. Singh and K. Kaur, "New Method for Finding Initial Cluster Centroids in K-means Algorithm," *Int. J. Comput. Appl.*, 2013, doi: 10.5120/12890-9837.

[13] Y. Li, J. Cai, H. Yang, J. Zhang, and X. Zhao, "A Novel Algorithm for Initial Cluster Center Selection," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2921320.