Progresif: Jurnal Ilmiah Komputer
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877
p-ISSN: 0216-3284

Analisis Perbandingan Pengukuran Jarak pada Algoritme *K-Means* Berbasis *Sum of Square Error*

Stendy Budi Hartono Sakur
Program studi Sistem Informasi, Politeknik Negeri Nusa Utara, Tahuna, Indonesia
e-mail: sakur.stendy@gmail.com

Abstract

The marketing strategy is very important to follow the culture of visitors or buyers because it is closely related to people's income levels. A number of visitor data are a data mining model that can extract information to determine the characteristics of each data. The purpose of this research is to compare distance measurements using the k-means clustering algorithm to see the optimal k value and the required time complexity. Using the K-Means clustering method with Euclidean, Manhattan, Minkowsky, Chebyshev, and Canberra distances to calculate the characteristic values of each object. Determining the value of k using the Elbow model which is formed from the Sum of Square Error (SSE) also considers the Mean of Square Error (MSE) value. The results showed that the Euclidean, Manhattan, Minkowsky, and Chebyshev distances can provide the right grouping so that they become an alternative to the Euclidean distance where the time needed by the Manhattan distance is 1.70 seconds faster than the Euclidean distance of 1.78 seconds, Minkowsky distance 1.82 seconds, Chebyshev distance 2.30 seconds and Canberra distance of 2.48 seconds. In conclusion, Euclidean, Manhattan, Minkowsky and Chebyshev distances can be used to measure closeness values between objects with good accuracy while Canberra distance cannot provide precise accuracy. The research resulted in five groups with different characteristics of income and expenses so that they can be used as a standard for developing marketing strategies.
Keywords: *K-means; Euclidean; Manhattan; Minkowsky; Chebyshev; Canberra; Sum of square error, Mean of square error*.

Abstrak

Strategi pemasaran sangat penting untuk mengikuti budaya pengunjung ataupun pembeli karena erat hubunganya dengan tingkat pendapatan masyarakat. Sejumlah data pengunjung merupakan suatu model data mining yang dapat digali informasinya guna mengetahui karakteristik dari setiap data. Tujuan penelitian ini adalah untuk membandingkan pengukuran jarak pada Algoritme *K-means clustering* sehingga diperoleh nilai k yang optimal serta kompleksitas waktu yang dibutuhkan. Menggunakan Metode *K-Means clustering* dengan *Euclidean, Manhattan, Minkowsky, Chebyshev* dan *Canberra distance* untuk menghitung nilai karakteristik dari setiap objek. Penentuan nilai k menggunakan kurva *Elbow* yang dibentuk dari *Sum of Square Error* (SSE) juga mempertimbangkan nilai *Mean of Square Error* (MSE). Hasil penelitian menunjukkan *Euclidean, Manattan, Minkowsky, Chebyshev distance* dapat memberikan pengelompokkan yang tepat sehingga menjadi alternatif pengganti *Euclidean distance* dimana waktu yang dibutuhkan oleh *Manhattan distance* sebesar 1.70 detik lebih cepat dibandingkan *Euclidean distance* 1.78 detik, *Minkowsky distance* 1.82 detik, *Chebyshev distance* 2.30 detik dan *Canberra distance* 2.48 detik. Kesimpulannya, *Euclidean, Manhattan, Minkowsky* dan *Chebyshev distance* dapat digunakan untuk mengukur nilai kedekatan antara objek dengan akurasi yang baik sedangkan *Canberra distance* tidak dapat memberikan akurasi dengan tepat. Penelitian menghasilkan 5 kelompok dengan karakteristik penghasilan dan pengeluaran yang berbeda sehingga dapat dijadikan sebagai standar pengembangan strategi pemasaran.

Kata Kunci**:** *K-means; Euclidean; Manhattan; Minkowsky; Chebyshev; Canberra; Sum of square error; Mean of square error*

1. Pendahuluan

Analisis *cluster* merupakan pembelajaran tidak terawasi merupakan teknik *multivariant* untuk mengeksplorasi data dengan menganalisis dan peringkasan data [1] yang lebih sulit dan menantang dibandingkan dengan klasifikasi, atau merupakan teknik klasifikasi statistik untuk menentukan posisi objek dari populasi kedalam kelompok tertentu dengan perbandingan kuantitatif dari beberapa karkateristik [2]. Metode *clustering* terdiri dari *Density-based method*, *Hieararhical based method, Partitioning method , Grid-based Methods* [3] dan *Model based Clustering* [4]. K-means merupakan *Partitioning method clustering* bersama dengan metode lainnya yaitu *K-Means, K-Medoids, K-Modes*, PAM, CLARANAS, CLARA, FCM, FCMdC, *Fanny* [4], *K-Protitype* [2] dan *K-Means++*. Proses pengelompokan data diperoleh dengan menerapkan pengukuran jarak pada setiap objek kemudian dengan algoritme tertentu dilakukan pengamatan kedekatan antar objek [4].

Terdapat beberapa pengukuran jarak yang dapat digunakan diantarnya *Euclidean distance, Canberra distance, Manhattan distance, Minkowsky distance, Chebyshev distance, Bit-vector distance, Hamming distance, Jaccard index, Cosine index, Dice Index* [5], [6] dan *Standard Euclidean distance, Mahalanobis distance, Standard Malahanobis distance* [7], *Bray curtis distance* [8]. Selain masalah jarak, masalah penentuan jumlah klaster (K) dan penentuan titik awal *centroid* dari *cluster* yang menjadi isu yang populer. Penentuan dan optimasi nilai K dapat dilakukan dengan menggunakan beberapa cara diantaranya dengan metode *Elbow*. Konsep dasar dari algoritme *elbow* adalah menggunakan jarak kuadrat dari setiap titik data dengan *centroid* untuk menghasilkan nilai K yang akan dihitung nilai *Error* dengan SSE, semakin kecil nilainya menunjukkan nilai yang konverge [9]. Selain itu penggunaan algoritme elbow dapat menggunakan *Within-Cluster Sum-of-Square* (WCSS)[10].

Penggunaan metode *clustering* dengan melakukan perbandingan jarak *Euclidean distance, Canberra distance* dan *Manhattan distance* telah dilakukan oleh Faisal dkk [6], dalam penelitian tersebut menggunakan *Z-Score* dan *Min-Max* Normalisasi dengan menggunakan koefisien *Silhouette*. Hasilnya menunjukkan *Canberra* memberikan hasil yang lebih baik pada *clustering* data iris. Suwanda [11] dalam penelitianya menunjukkan *Manhattan distance* memberikan hasil yang maksimal dibandingkan dengan *Euclidean distance* ketika melakukan perbandingan berdasarkan variasi jumlah klaster (k) dengan menggunakan sampel data dari repositori UCI *machine learning*.

Ardianti, melakukan penelitian dengan metode K-NN dan *Color Extraction method* yang menggunakan *Manhattan distance* dan *Euclidean distance* [12]. Sakur [13] melakukan perbandingan jarak *Euclidean distance, Manhattan distance* dan *Minkowsky distance* dengan algoritme *K-means* untuk proses pengelompokan dan menggunakan metode TOPSIS untuk proses perangkingan dengan menggunakan korelasi *Pearson* dan *Spearman*. Sakur [14] melakukan perbandingan jarak *Euclidean, Manhattan* dan *Minkowsky* dengan cara yang berbeda yaitu mengitung klaster dan perangkingan setiap jarak yang digunakan.

Beberapa penelitian yang berfokus pada optimalisasi pencarian nilai K seperti pada literatur review ashabi [15] yang menguraikan berbagai penelitian dalam mengoptimasliasi nilai K. Cui yang menggunakan metode Elbow dengan WCSS[10], Ekasetya menggunakan nilai *Sum of Square Error* (SSE) untuk menentukan nilai optimal dari K [16], Pugazhenthi, meneliti sejumlah parameter dari *Peak signal to Noise ratio, structural content, Means Square Error*, *Stuctural similarity index, Universal Quality Index, Correlation Coefficient* dan *Image Fidality* yang dapat meningkatkan optimalisasi penentuan nilai *centroid* secara acak [17], yuan melakukan penelitian untuk penentuan nilai K yang terdiri dari empat algoritme yaitu *Elbow Method*[18], *Gap Statistic, Silhoutte Coefficient* dan *Canopy* untuk mengklaster data Iris dengan hasil k=2, waktu eksekusi 1.830 s, 9.763ss, 8.648ss, 2.120 s secara berurutan dari *Elbow Method, Gap Statistic, Silhoutte Coefficient* dan *Canopy* [9]. Untuk validasi klaster dapat digunakan *Purity* dan *Deavies-Bouldin Index* (DBI)[8], *Bayesian information criteria* (BIC), *Akaike information criterian* (AIC), *Dunn's index, Silhouette width* (SW), *Calinski and Harabasz index* (CH), *Gap Static, Generalized Dunn's index* (DNg) dan *Modified Dunn's index* [19]. Penelitian dari Dista [20] bertujuan untuk melakukan optimasi dengan *Particle Swarm Optimaztion* (PSO) untuk menentukan seberapa baik *cluster* yang dihasilkan namun waktu yang dibutuhkan lebih lama.

Dari penelitian terdahulu terlihat bahwa pengukuran jarak dengan proses optimasi nilai K memiliki pengaruh terhadap hasil *cluster* yang dibentuk. Beberapa penelitian hanya menitikberatkan pada jarak *euclidean, manhattan dan minkowsky* [2], [12], [14] akan tetapi

penting untuk menyelidikan jarak lainnya seperti jarak *Chebyshev* ataupun *Canberra* yang juga banyak digunakan pada proses *clustering*. Selain itu, proses optimalisasi yang telah dilakukan oleh peneliti [20] merupakan model optimasi terhadap hasil dari clustering dengan waktu eksekusi yang digunakan cukup lama sehingga penting menampilkan waktu eksekusi dari penelitian ini untuk menunjukkan hasil cluster yang sama dengan waktu eksekusi yang lebih cepat.

Tujuan penelitian adalah melakukan perbandingan pengukuran jarak dari *Euclidean distance, Manhattan distance, Minkowsky distance, Chebyshev distance* dan *Canberra distance* dengan menentukan nilai K optimal berdasarkan pada nilai *Sum of Square Error* (SSE) dengan memperhitungkan waktu eksekusi yang dilakukan.

2. Metodologi

Clustering digunakan untuk mengelompokkan berdasarkan tingkat kemiripan objek. Dengan menggunakan algoritme *K-Means* akan dilakukan pengujian terhadap dataset yang ada. Penelitian ini menggunakan dataset dari Kaggle [23] yaitu Dataset pengunjung mall, yang terdiri dari 200 *customer* dengan kategori Gender, Usia, Annual income dan Spending score dalam skala 1 – 100. Lingkungan pengujian dilakukan pada *device* Laptop dengan spesifikasi Intel Core i7 degan memory 16 GB, dengan operating sistem *Linux Deepin* Versi 20. Untuk mempercepat proses iterasi pada *cluster* maka pengembangan sistem dibuat dengan bahasa pemrograman Qt/C++ Framework [24], dengan menggunakan pustaka Aljabar Liniear Armadillo C++ [25] menggunakan tipe data *Spares Matrix* dari Armadillo [26]. Dataset awal berbentuk file .CSV yang kemudian diload kedalam program untuk ditamplkan pada tabel, kemudian dikonversikan kedalam Spares matrix untuk kategori *Annual income* dan *Spending score* (1-100). Penentuan nilai awal centroid dilakukan secara acak dan manual berdasarkan nilai K yang dimulai dari 2,3,4,5,6,7,8,9,10. Masing-masing nilai K akan menjalankan algoritme *K-Means* dengan menggunakan lima jarak yang berbeda yaitu *Euclidean, Manhattan, Minkowsky, Chebyshev* dan *Canberra*. Setiap iterasi yang dilakukan sistem akan menghitung nilai SSE dan MSE untuk kemudian disimpan dalam file .TXT yang akan di proses akhir pada aplikasi *Libreoffice Calc* atau juga bisa pada aplikasi *Excel*. Secara umum algoritme *K-Means* sebagai berikut:

1. Step 1: Penentuan nilai k yang terdiri dari 2,3,4,5,6,7,8,9,10.
2. Step 2: menentukan nilai centroid awal sesuai dengan jumlah k, secara acak dan manual,
3. Step 3: hitung jarak awal setiap titik data dengan jumlah *centroid* yang digunakan menggunakan persamaan,

Euclidean distance [2], [6], [13], [27], [28] ,

$$\boldsymbol{dist_{ij}} = \sqrt{\sum_{k=1}^{n}\left(\boldsymbol{x_{ik}} - \boldsymbol{y_{jk}}\right)^2} \quad (1)$$

Manhattan distance (City block) [2], [6], [27],

$$\boldsymbol{dist_{ij}} = \sum_{k=1}^{n}\left|\boldsymbol{x_{ik}} - \boldsymbol{y_{jk}}\right| \quad (2)$$

Minkowsky distance [2],

$$\boldsymbol{dist_{ij}} = \left(\sum_{k=1}^{n}\left|\boldsymbol{x_{ik}} - \boldsymbol{y_{jk}}\right|^{p}\right)^{\frac{1}{p}} \quad (3)$$

Nilai p pada penelitian ini diambil 4, agar tidak kembali seperti pada *Eculidean, Manhattan* ataupun *Chebysev*.

Chebyshev distance [2],

$$\boldsymbol{dist_{ij}} = \left|\boldsymbol{x_{ik}} - \boldsymbol{y_{jk}}\right| \quad (4)$$

Canberra distance [6], [8],

$$dist_{ij} = \sqrt{\sum_{k=1}^{n} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| - |x_{jk}|}} \quad (5)$$

4. Step 4: tentukan nilai terdekat dengan *centroid* dan tentukan lokasi class dari data.
5. Step 5: hitung nilai *Sum of Square Error* (SSE) dari setiap kelas yang terbentuk dengan persamaan berikut [29],

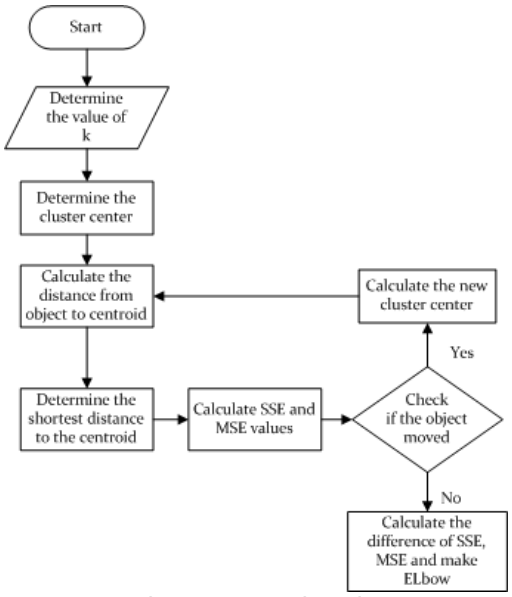
$$\mathbf{SSE} = \sum_{k=1}^{n} \sum_{\forall x_i \epsilon C_k} \|x_i - \overline{x_k}\|^2 \quad (6)$$

6. Step 6: hitung nilai *Mean of Square Error* (MSE) dari hasil perhitungan SSE pada Step 5, dengan persamaan berikut [30],

$$\mathbf{MSE} = \frac{1}{n} \sum_{k=1}^{n} \sum_{\forall x_i \epsilon C_k} \|x_i - \overline{x_k}\|^2 \quad (7)$$

7. Step 7: periksa apakah masih ada perubahan klaster, jika masih ada perubahan *cluster* maka lakukan Step 2 – Step6.
8. Step 8: diassumsikan sudah tidak ada perubahan klaster atau iterasi telah konvergen maka nilai SSE dan MSE terakhir diambil untuk digunakan pada pembuatan algoritme *Elbow*.
9. Step 9: proses selesai.

Dalam perhitungan *cluster*, dilakukan penyimpanan data hasil perhitungan SSE dan MSE termasuk waktu yang dibutuhkan untuk menyelesaikan setiap iterasi pada setiap nilai k. Perhitungan waktu menggunakan fungsi dari Qt/C++ yaitu header *QElapsedTimer*. Setelah proses perhitungan selesai maka data akan diolah pada aplikasi *Excel*. Gambar 1, menunjukkan proses perhitungan *K-means* dimana proses tersebut akan dilakukan sejumlah pengukuran jarak yang digunakan dengan nilai k dari 2 hingga 10.



Gambar 1. Prosedur Klaster

3. Hasil dan Pembahasan

3.1. Hasil pengujian

Dataset diekstrak untuk mengambil kolom yang digunakan. Tabel 1, merupakan hasil ekstrasi untuk menyimpan kolom. Kemudian tentukan nilai k secara manual dan dilanjukan dengan proses *cluster* menggunakan *Eculidean, Manhattan, Minkowsky, Chebyshev* dan *Canberra distance*. Dengan menggunakan model *Thread* dari Qt/C++ proses *clustering*

dilakukan sehingga tidak mengganggu sistem utama. Hasil perhitungan memberikan nilai SSE dan MSE yang membentuk kurva *Elbow* untuk menentukan nilai k yang optimal.

Tabel 1. Dataset yang sudah diekstrak

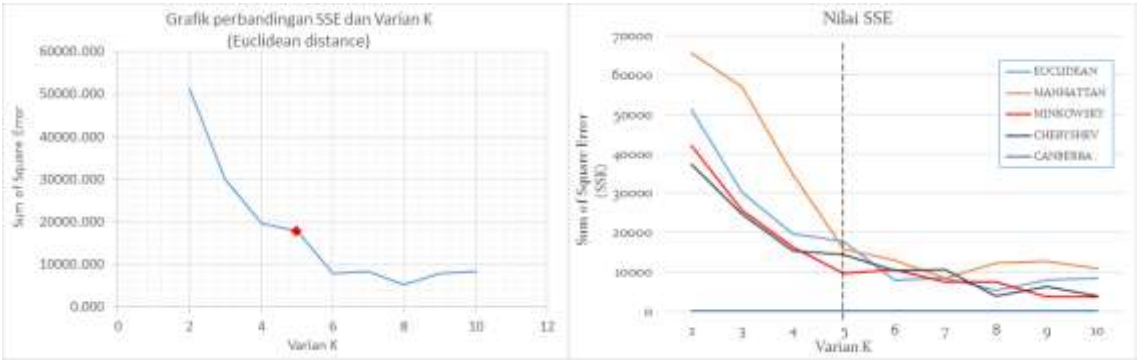
Annual Income (k$)	*Spending Score* (1-100)
15	39
15	81
16	6
16	77
17	40
…	…
196	16
197	28
198	74
199	18
200	83

Tabel 2, merupakan hasil perhitungan nilai SSE beserta selisih nilai untuk perhitungan kurva *Elbow* dengan *Euclidean distance*.

Tabel 2. Nilai SSE yang terbentuk

K	SSE	Selisih
2	51348,700	0
3	30178,800	21170
4	19715,300	10464
5	17895,600	1820
6	7899,060	9997
7	8402,320	503
8	5414,330	2988
9	7902,810	2488
10	8442,080	539

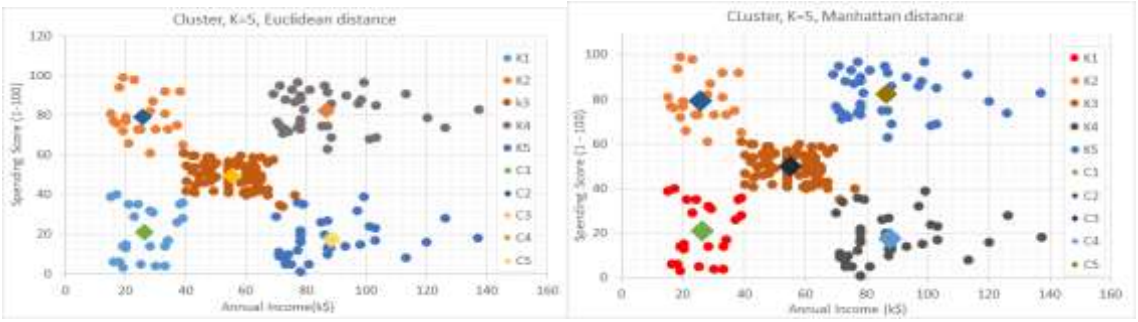
Gambar 3 (kiri) menunjukkan sebuah perubahan bentuk kurva *Elbow* namun masih belum sempurna, berbeda halnya dengan bentuk *Elbow* dari *Manhattan distance*. Gambar 3 (kanan), menunjukkan penggabungan dari seluruh nilai SSE untuk seluruh jarak yang digunakan,



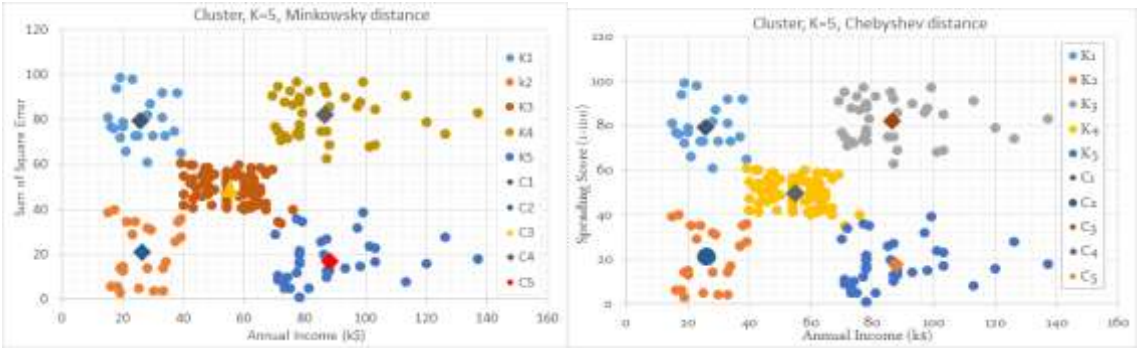
Gambar 3. Nilai SSE untuk *Euclidean distance* (kiri) dan Gabungan nilai SSE *dari Euclidean, Manhattan, Minkowsky, Chebyshev* dan *Canberra distance* (kanan).

Gambar 3 (posisi kanan), menunjukkan nilai k = 5 sebagai titik optimal untuk *cluster*. Sekalipun pembentukan model *Elbow* belum mencapai bentuk yang sempurna. Dari kelima jarak tersebut bentuk yang mendekati sempurna adalah *Manhattan, Chebyshev* dan *Minkowsky distance*. Gambar 5 - 7 merupakan hasil *cluster* dengan k=5 untuk semua jarak yang digunakan, dimana setiap jarak membentuk posisi *cluster* sendiri oleh karena itu perlu

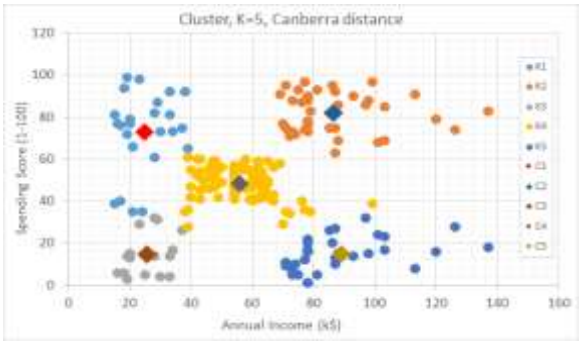
diseragamkan agar mudah dianalisis. Gambar 8, posisi *cluster* yang baru sesuai hasil dari *Euclidean* dan *Manhattan distance*. Tabel 3 terdiri dari total data dari setiap klaster sesuai posisi *cluster* dari Gambar 8 dan data point sesuai Gambar 5 hingga 7. Gambar 9 (kiri), menunjukkan perbedaan jumlah iterasi untuk setiap nilai k dan jarak yang digunakan.



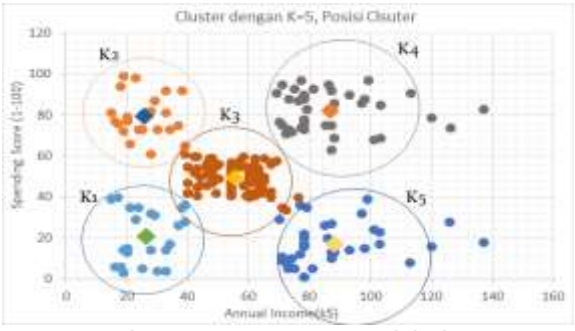
Gambar 5. Klaster k=5 dari *Eculidean distance* (kiri) dan *Manhattan distance* (kanan)



Gambar 6. Klaster k=5 dari *Minkowsky distance* (kiri) dan *Chebyshev distance* (kanan)



Gambar 7. Klaster k=5 dari *Canberra distance*

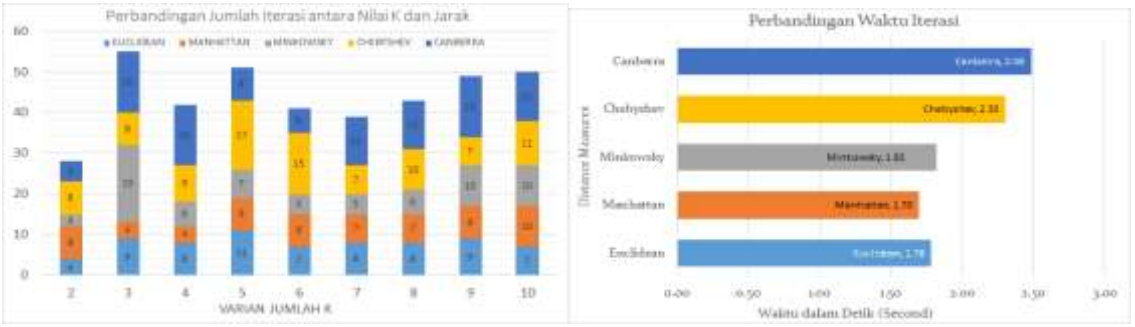


Gambar 8. Pengaturan posisi *cluster*

Tabel 3. Jumlah data dalam *cluster*

Jarak	*Cluster*				
	1	2	3	4	5
Euclidean	23	22	81	39	35
Manhattan	23	22	80	39	36
Minkowsky	23	22	81	39	35
Chebyshev	23	22	80	39	35
Canberra	16	26	88	39	31

Untuk mengukur kompleksitas waktu iterasi, digunakan fungsi *QElapsedTimer* dari Framework Qt/C++ yang dihitung mulai awal iterasi dimana *QThread* dijalankan. Gambar 9 (kanan) merupakan perbandingan waktu eksekusi yang diambil rata-rata dari k=2 sampai dengan k=10 pada setiap pengukuran jarak.



Gambar 9. Perbandingan Hasil iterasi (kiri) dan Waktu eksekusi program (kanan)

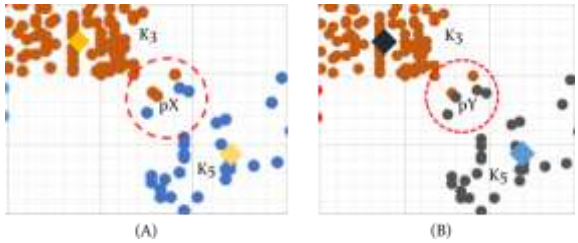
3.2. Pembahasan

Dari hasil pengujian yang dilakukan terdapat beberapa bagian penting yang perlu dibahas yaitu, 1) Untuk mendapatkan nilai k yang optimal maka perlu dilakukan pengujian berulang kali. Dengan 2x pengujian kurva *Elbow* berdasarkan SSE mendapatkan nilai yang cukup jauh berbeda pada posisi k =3, 4 dan 5 untuk *Euclidean distance* dan *Chebyshev diatance*. *Manhattan* dan *Minkowsky distance* secara visual terlihat dengan jelas perubahan bentuk pada posisi k=5. Dengan demikian secara keseluruhan k yang optimal berada pada posisi 5. Jika dibandingkan dengan penelitian dari [20] nilai *Elbow* pada jarak *euclidean* memiliki sedikit berbeda namun nilai siku yang tepat berada pada posisi 5 hal ini diperkuat dengan jarak *manhattan, minkowsky* dan *canberra* (untuk *canberra* dapat dilihat pada Gambar 10). 2) *Cluster* yang terbentuk secara umum memiliki kesamaan dari seluruh jarak. *Cluster* yang terbentuk secara visual terlihat pda Gambar 5 hingga 7, dimana *Canberra distance* memiliki *cluster* yang tidak baik (khusus penelitian ini) dimana terdapat objek yang seharusnya masuk sebagai *cluster* 1 namun masuk ke *cluster* 2, demikian pula terdapat objek yang seharusnya masuk dalam *cluster* 4 namun berada pada posisi *cluster* 3 sehingga dengan pengujian yang minim terlihat *canberra* tidak dapat melakukan pengelompokkan dengan tepat.

Gambar 10. Nilai SSE dan k optimal

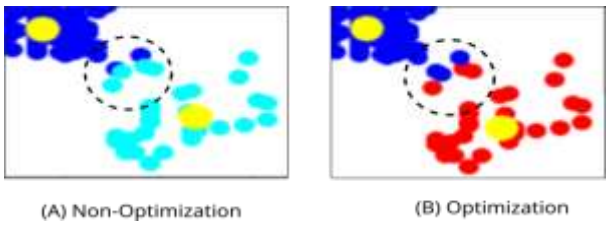
Dari Tabel 3 dan Gambar 5 hingga 7 terlihat *cluster* dengan jarak *Euclidean distance* dan *Minkowsky distance* memberikan hasil *clustering* yang tepat sama, demikian juga *Manhattan distance* dan *Chebyshev distance* memberikan hasil yang sama pula sedangkan untuk

Canberra perlu diselidiki lebih lanjut untuk ketepatan nilai *cluster*. Perbedaan *cluster* dari group jarak *Eculidean* dan *Minkowsky dengan Manhattan* dan *Chebyshev* terlihat pada Gambar 11,



Gambar 11. Perbedaan *cluster* dari Group X (*Eculidean, Minkowsky*) dan Group Y (*Mahattan, Chebyshev*)

Gambar 11, menunjukkan perbedaan dimana titik pX pada Gambar 11 (A) masuk kedalam *cluster* K3 sedangkan untuk titik pY pada Gambar 11 (B) masuk kedalam *cluster* K5, sehingga dibutuhkan pembanding lain untuk melihat akurasi *cluster*. Namun dari hasil *cluster* diatas dapat dilihat bahwa keempat pengukuran jarak dapat digunakan secara akurat untuk proses pengelompokan data. Dimana hasil dari *Euclidean* dan *Minkowsky* memiliki hasil yang sama begitu juga untuk *Manhattan* dan *Chebyshev*. Sebagai perbandingan maka peneliti menggunakan penelitian [20] yang menggunakan *Particle Swarm Optimization* (PSO) untuk mengoptimalkan hasil *clustering*, seperti terlihat pada Gambar 12,



Gambar 12. Hasil perbandingan *Cluster* sebelum dan sesudah optimasi menggunakan *Particle Swarm Optimization* (PSO) berdasarkan penelitian *[20]*

Gambar 12 merupakan perbedaan cluster dimana (A) sebelum dioptimasi dan (B) sesudah dioptimasi dengan *Euclidean distance*. Jika diambil perbandingan maka *cluster* yang belum dioptimasi memiliki hasil sama dengan *cluster* menggunakan *Manhattan distance* dan *Chebyshev distance*, sedangkan *cluster* yang telah dioptimasi, Gambar 12 (B) sama dengan *Euclidean distance* dan *Minkowsky distance*. Sehingga keempat jarak tersebut dapat digunakan sebagai alternatif dari *euclidean distance*. 3) jika dilihat dari jumlah iterasi maka *Manhattan distance* memberikan hasil yang baik karena memiliki rata-rata jumlah iterasi sebesar 7.11, sedangkan iterasi terbanyak ada pada *Canberra distance*. 4) dari kompleksitas waktu konvergen maka *Manhattan distance* membutuhkan waktu 1.70 detik, lebih cepat dibandingkan dengan *Eculidean distance* 1,78 detik, *Minkowsky distance* dengan 1.82 detik, *Chebyshev distance* dengan 2.30 detik dan *Canberra distance* 2.48 detik. Dari hasil experimen yang dilakukan dapat dijelaskan hasil akhir dari dataset pengunjung mall dengan merujuk posisi *cluster* dan hasilnya pada Gambar 8 dan Gambar 5 hingga 7, yaitu:

1) Kelompok 1: merupakan pengunjung mall yang memiliki penghasilan yang rendah berkisar $10 hingga $40 dengan kemampuan pengeluaran yang juga rendah sehingga pihak mall tidak harus memfokuskan pada kelompok ini.
2) Kelompok 2: merupakan kelompok dengan tingkat pengeluaran yang tinggi tapi memiliki penghasilan yang rendah diantara $10 hingga $40, kelompok ini cenderung atau gemar berbelanja sekalipun penghasilan yang tidak berimbang.
3) Kelompok 3: merupakan kelompok dengan penghasilan yang sedang dan pengeluaran yang sedang juga bisa dikatakan bahwa kelompok ini selalu menyeimbangkan kebutuhan dan pendapatannya.
4) Kelompok 4: merupakan kelompok yang menjadi target dari pengelola mall karena memiliki pendapatan yang tertinggi berkisar $70 hingga $140 dengan tingkat

pengeluaran yang tinggi. Pengelola mall harus memiliki strategi yang baik untuk mempertahankan pangsa pasar ini.

5) Kelompok 5: merupakan kelompok yang strategis juga karena memiliki penghasilan yang sama dengan klaster 4 namun tingkat pengeluaran yang sedikit atau kecil. Pengelola mall harus memiliki strategi yang jitu untuk menarik keinginan dari kelompok ini agar dapat memliki keinginan untuk berbelanja.

4. Simpulan

Hasil penelitian menunjukkan bahwa *Euclidean distance, Manhattan distance, Minkowsky distance* dan *Chebyshev distance* memberikan hasil yang akurat untuk *cluster* yang dibentuk dengan perbandingan dari penelitian sebelumnya. Sedangkan untuk *Canberra distance* memiliki *cluster* yang kurang tepat karena banyak kelompok *cluster* yang bersilangan sehingga perlu diselidiki lebih lanjut lagi. *Manhattan distance* memiliki waktu eksekusi yang lebih kecil dari keempat jarak lainnya sehingga dapat dijadikan alternatif pengganti dari *euclidean distance*. Penelitian menghasilkan 5 kelompok yang akurat dan dapat memberikan unjuk kerja yang lebih baik dari penelitian sebelumnya. Pengelompokan yang terbentuk dapat menentukan karakteristik dari setiap pelanggan sehingga dapat membantu pihak pengelola untuk membuat strategi marketing yang lebih baik lagi.

Daftar Referensi

[1] V. V. Hegde and N. S. Gadwal, “A Review on Cloud Computing and K-means++ Clustering Algorithm with Map Reduce,” *IJRESM*, vol. 2, no. 5, pp. 526–528, May 2019.

[2] R. Nooraeni and G. Nurfalah, “Kajian Penerapan Jarak Euclidean, Manhattan, Minkowski, dan Chebyshev pada Algoritme Clustering K-Prototype,” vol. 4, no. 2, Art. no. 2, 2022.

[3] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, “Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means,” in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India: IEEE, Mar. 2020, pp. 306–310. doi: 10.1109/ICCMC48092.2020.ICCMC-00057.

[4] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, “A Short Review on Different Clustering Techniques and Their Applications,” in *Emerging Technology in Modelling and Graphics*, J. K. Mandal and D. Bhattacharya, Eds., in Advances in Intelligent Systems and Computing, vol. 937. Singapore: Springer Singapore, 2020, pp. 69–83. doi: 10.1007/978-981-13-7403-6_9.

[5] S. Pandit and S. Gupta, “A Comparative Study on Distance Measuring Approaches for Clustering,” *IJORCS*, vol. 2, no. 1, Art. no. 1, Dec. 2011, doi: 10.7815/ijorcs.21.2011.011.

[6] M. Faisal, E. M. Zamzami, and Sutarman, “Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance,” *J. Phys.: Conf. Ser.*, vol. 1566, no. 1, Art. no. 1, Jun. 2020, doi: 10.1088/1742-6596/1566/1/012112.

[7] J. Arora, K. Khatter, and M. Tushir, “Fuzzy c-Means Clustering Strategies: A Review of Distance Measures,” in *Software Engineering*, Springer, Singapore, 2019, pp. 153–162. doi: 10.1007/978-981-10-8848-3_15.

[8] F. A. Sebayang, M. S. Lydia, and B. B. Nasution, “Optimization on Purity K-Means Using Variant Distance Measure,” in *2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)*, Medan, Indonesia: IEEE, Jun. 2020, pp. 143–147. doi: 10.1109/MECnIT48290.2020.9166600.

[9] C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” *J*, vol. 2, no. 2, Art. no. 2, Jun. 2019, doi: 10.3390/j2020016.

[10] M. Cui, “Introduction to the K-Means Clustering Algorithm Based on the Elbow Method,” *Accounting, Auditing and Finance*, vol. 1, no. 1, Art. no. 1, Oct. 2020, doi: 10.23977/accaf.2020.010102.

[11] R. Suwanda, Z. Syahputra, and E. M. Zamzami, “Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K,” *J. Phys.: Conf. Ser.*, vol. 1566, no. 1, Art. no. 1, Jun. 2020, doi: 10.1088/1742-6596/1566/1/012058.

[12] W. F. Ardianto, S. Sriani, and A. H. Hasugian, “Application of color extraction methods and k-nearest neighbor to determine maturity avocado butter,” *J. Teknik Informatika CIT Medicom,* vol. 15, no. 1, Art. no. 1, Mar. 2023, doi: 10.35335/cit.Vol15.2023.375.pp09-20.

[13] S. B. H. Sakur, “PERBANDINGAN DISTANCE MEASURES PADA K-MEANS CLUSTER DAN TOPSIS DENGAN KORELASI PEARSON DAN SPEARMAN,” *JITEK*, vol. 3, no. 1, Art. no. 1, Mar. 2023, doi: 10.55606/jitek.v3i1.1394.
[14] S. B. H. Sakur, M. Silangen, and D. Tuwohingide, “Penerapan Algoritme K-Means Cluster dan Metode TOPSIS pada Pemilihan Mahasiswa kunjungan Industri,” *Jutisi : Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, vol. 11, no. 3, Art. no. 3, Dec. 2022, doi: 10.35889/jutisi.v11i3.1045.
[15] A. Ashabi, S. B. Sahibuddin, and M. Salkhordeh Haghighi, “The Systematic Review of K-Means Clustering Algorithm,” in *2020 The 9th International Conference on Networks, Communication and Computing*, Tokyo Japan: ACM, Dec. 2020, pp. 13–18. doi: 10.1145/3447654.3447657.
[16] V. A. Ekasetya and A. Jananto, “KLASTERISASI OPTIMAL DENGAN ELBOW METHOD UNTUK PENGELOMPOKAN DATA KECELAKAAN LALU LINTAS DI KOTA SEMARANG,” *JDI*, vol. 12, no. 1, Art. no. 1, Aug. 2020, doi: 10.35315/informatika.v12i1.8159.
[17] A. Pugazhenthi and L. S. Kumar, “Selection of Optimal Number of Clusters and Centroids for K-means and Fuzzy C-means Clustering: A Review,” in *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, Patna, India: IEEE, Oct. 2020, pp. 1–4. doi: 10.1109/ICCCS49678.2020.9276978.
[18] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster,” *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 336, no. 1, Art. no. 1, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.
[19] K. P. Sinaga and M.-S. Yang, “Unsupervised K-Means Clustering Algorithm,” *IEEE Transactions on Fuzzy Systems*, vol. 8, pp. 80716–80727, May 2023, doi: 10.1109/ACCESS.2020.2988796.
[20] T. M. Dista and F. F. Abdulloh, “Clustering Pengunjung Mall Menggunakan Metode K-Means dan Particle Swarm Optimization,” *mib*, vol. 6, no. 3, Art. no. 3, Jul. 2022, doi: 10.30865/mib.v6i3.4172.
[21] R. Nainggolan and G. Lumbantoruan, “OPTIMASI PERFORMA CLUSTER K-MEANS MENGGUNAKAN SUM OF SQUARED ERROR (SSE),” vol. 2, no. 2, Art. no. 2, 2018.
[22] “Perbaikan Kinerja Clustering K-Means pada Data Ekonomi Nelayan dengan Perhitungan Sum of Square Error (SSE) dan Optimasi nilai K cluster,” *ResearchGate*, Mar. 2023, doi: 10.33633/tc.v20i2.4572.
[23] “Data-Pengunjung-Mall,” May 30, 2023. https://www.kaggle.com/datasets/baktisiregar/datapengunjungmall (accessed May 30, 2023).
[24] “Qt Framework: C++/Python/QML,” *web official Qt Framework*, May 30, 2023. https://www.qt.io/product/framework (accessed May 30, 2023).
[25] C. Sanderson and R. Curtin, “Armadillo: a template-based C++ library for linear algebra,” *JOSS*, vol. 1, no. 2, Art. no. 2, Jun. 2016, doi: 10.21105/joss.00026.
[26] C. Sanderson and R. Curtin, “A User-Friendly Hybrid Sparse Matrix Class in C++,” in *Mathematical Software – ICMS 2018*, J. H. Davenport, M. Kauers, G. Labahn, and J. Urban, Eds., in Lecture Notes in Computer Science, vol. 10931. Cham: Springer International Publishing, 2018, pp. 422–430. doi: 10.1007/978-3-319-96418-8_50.
[27] S. M. H. M. Huzir, N. Z. Mahabob, A. F. M. Amidon, N. Ismail, Z. M. Yusoff, and M. N. Taib, “A Ppreliminary study on the intelligent model of k-nearest neighbor for agarwood oil quality grading,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 3, Art. no. 3, Sep. 2022.
[28] R. L. Lafta, M. S. AL-Musaylh, and Q. M. Shallal, “Clustering similar time series data for the prediction the patients with heart disease,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, Art. no. 2, May 2022.
[29] A. Smiti, “A critical overview of outlier detection methods,” *Computer Science Review*, vol. 38, p. 100306, Nov. 2020, doi: 10.1016/j.cosrev.2020.100306.
[30] S. N. Wahyuni, E. Sediono, I. Sembiring, and N. N. Khanom, “Comparative analysis of time series prediction model for forecasting COVID-19 trend,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, Art. no. 1, Oct. 2022.