Jurnal Informatika: Jurnal Pengembangan IT (JPIT), Vol.04, No.01, Januari 2019 ISSN: 2477-5126
DOI: 10.30591/jpit.v4i1.1253 e-ISSN: 2548-9356

Perbandingan Akurasi *Euclidean Distance*, *Minkowski Distance*, dan *Manhattan Distance* pada Algoritma *K-Means Clustering* berbasis *Chi-Square*

M. Nishom[*)]
Jurusan Teknik Informatika, Politeknik Harapan Bersama, Tegal
Jl. Mataram No.9 Pesurungan Lor Kota Tegal, 52147, Indonesia
email: nishom@poltektegal.ac.id

***Abstract* – In data mining, there are several algorithms that are often used in grouping data, including K-Means. However, this method still has several disadvantages, including the problem of the level of accuracy of the methods used to measure the similarities between the objects being compared. To overcome this problem, in this study a comparison was made between three methods (euclidean distance, manhattan distance, and minkowski distance) to determine the status of disparity in Teacher's needs in Tegal City. The results showed that of the three methods compared had a good level of accuracy, which is 84.47% (for euclidean distance), 83.85% (for manhattan distance), and 83.85% (for minkowski distance). In addition, this study also informs that there are still 6 (six) schools with conditions that are very poorly available for teachers (in the category of HIGH disparity labels) and need to get more attention, which is SMP Atmaja Wacana, SMKN 3 Tegal, SMAS Muhammadiyah, SMAS Pancasakti Tegal, SMKS Muhammadiyah 1 Kota Tegal, dan SMP IC Bias Assalam.**

***Abstrak* - Dalam data mining, terdapat beberapa algoritma yang sering digunakan dalam pengelompokan data, diantaranya adalah K-Means. Namun, metode tersebut masih memiliki beberapa kekurangan, diantaranya adalah masalah tingkat akurasi dari metode yang digunakan untuk mengukur kesamaan antar objek-objek yang dibandingkan. Untuk mengatasi permasalahan tersebut, dalam penelitian ini dilakukan komparasi antar tiga metode (euclidean distance, manhattan distance, dan minkowski distance) untuk mengetahui status disparitas kebutuhan Guru di Kota Tegal. Dataset yang digunakan dalam penelitian ini adalah data pokok pendidikan tingkat dasar dan menengah di Kota Tegal. Hasil penelitian menunjukkan bahwa dari ketiga metode yang dibandingkaan memiliki tingkat akurasi yang baik, yaitu 84.47% (untuk euclidean distance), 83.85% (untuk manhattan distance), dan 83.85% (untuk minkowski distance). Selain itu, dalam penelitian ini juga menginfomasikan bahwa masih tedapat 6 (enam) sekolah dengan kondisi ketersediaan Guru yang masih sangat kurang (masuk kategori label disparitas TINGGI) dan perlu mendapatkan perhatian lebih yaitu SMP Atmaja Wacana, SMKN 3 Tegal, SMAS Muhammadiyah, SMAS Pancasakti Tegal, SMKS Muhammadiyah 1 Kota Tegal, dan SMP IC Bias Assalam.**

Kata Kunci: euclidean distance, manhattan distance, minkowski distance, K-Means, disparitas kebutuhan guru

*) **Corresponding author:** (M. Nishom)
Email: nishom@poltektegal.ac.id

I. PENDAHULUAN

Clustering merupakan aktivitas (*task*) yang bertujuan mengelompokkan data yang memiliki kemiripan antara satu data dengan data lainnya ke dalam klaster atau kelompok sehingga data dalam satu klaster memiliki tingkat kemiripan (*similiarity*) yang maksimum dan data antar klaster memiliki kemiripan yang minimum. *Clustering* juga dapat diartikan metode segmentasi data yang diimplementasikan dalam beberapa bidang, diantaranya marketing, analisa masalah bisnis segmentasi pasar dan prediksi, pola dalam bidang computer vision, zonasi wilayah hingga identifikasi obyek dan pengolahan citra. Analisis klaster bertujuan menemukan kelompok objek sedemikian rupa sehingga objek-objek dalam grup akan sama (atau terkait) satu sama lain dan berbeda dari (atau tidak terkait) objek-objek dalam grup lain [1]. Ada sejumlah algoritma yang dapat digunakan untuk pengelompokan. Secara umum, *K-Means* merupakan algoritma heuristik yang memisah kumpulan data ke dalam klaster K dengan meminimalkan jumlah jarak kuadrat di setiap klaster. Pada penelitian sebelumnya [2], penerapan algoritma *K-Means* dasar telah dilakukan dengan menggunakan metode pengukuran jarak *Euclidean* (*Euclidean Distance*). Pada penelitian ini, diimplementasikan pengukuran jarak menggunakan metode *Manhattan* dan *Minkowski* pada algoritma *K-Means* berbasis *Chi-Square*. Selain itu, juga dilakukan perbandingan tingkat akurasi dari masing-masing metode untuk mengetahui metode terbaik.

II. PENELITIAN YANG TERKAIT

Terdapat beberapa penelitian yang telah dilakukan sebelumnya menyebutkan bahwa *clustering* menggunakan algoritma k-means relatif lebih cepat dibandingkan *clustering* menggunakan algoritma lain, selain itu juga menghasilkan klaster yang berkualitas ketika menggunakan *dataset* berukuran besar [3]. Perbandingan metode perhitungan *Manhattan* dan *Euclidean distance* pada algoritma *k-means* untuk mengetahui jumlah *squared error* menggunakan *Bank dataset* dan diuji menggunakan tool WEKA. Dari hasil pengujian menunjukkan bahwa metode *Manhattan distance* lebih baik dari pada metode *Euclidean* [4]. Perbandingan 3 (tiga) metode perhitungan *distance* pada algoritma k-means (*Manhattan*, *Euclidean* dan *Minkowski)* untuk menemukan metode perhitungan jarak yang paling baik, hasil penelitian

menyimpulkan bahwa metode perhitungan jarak *Euclidean* lebih baik dari pada metode *Manhattan* dan *Minkowski* [5]. Penelitian lain tentang perbandingan metode perhitungan *Manhattan*, *Euclidean* dan *Chebyshev Distance* pada algoritma k-means untuk mengetahui *mean absolute error*. Hasil pengujian yang dilakukan menggunakan *flower dataset* menunjukkan bahwa metode perhitungan *Chebyshev Distance* lebih baik dari metode *Manhattan* dan *Euclidean*[6]. Pada penelitian yang berbeda, diketahui bahwa metode perhitungan jarak Manhattan, Euclidean dan *Chebyshev* saling unggul antara satu dengan yang lain tergantung data-set yang digunakan [7], [8], [9].

III. METODE PENELITIAN

Pengukuran jarak memegang peran yang sangat penting dalam menentukan kemiripan atau keteraturan di antara data dan item. hal ini dilakukan untuk mengetahui, dengan cara seperti apa data dikatakan saling terkait, mirip, tidak mirip, dan metode pengukuran jarak seperti apa yang diperlukan untuk membandingkannya [10]. Pada proses *clustering*, tahapan menentukan atau mendeskripsikan nilai kuantitatif dari tingkat kemiripan atau ketidakmiripan data (*proximity measure*) memiliki peranan sangat penting, sehingga perlu dilakukannya perbandingan beberapa metode yang sering digunakan, yaitu jarak *euclidean*, *manhattan*, dan *minkowski*.

A. *Euclidean Distance*

Euclidean distance merupakan salah satu metode perhitungan jarak yang digunakan untuk mengukur jarak dari 2 (dua) buah titik dalam *Euclidean space* (meliputi bidang *euclidean* dua dimensi, tiga dimensi, atau bahkan lebih). Untuk mengukur tingkat kemiripan data dengan rumus *euclidean distance* digunakan rumus berikut [11]:

$$d(x,y) = |x-y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

dimana,
d = jarak antara *x dan* y
x = data pusat klaster
y = data pada atribut
i = setiap data
n = jumlah data,
x_i = data pada pusat klaster ke *i*
y_i = data pada setiap data ke *i*

B. *Manhattan Distance*

Manhattan distance digunakan untuk menghitung perbedaan absolut (mutlak) antara koordinat sepasang objek. Rumus yang digunakan adalah sebagai berikut:

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (2)$$

dimana,
d = jarak antara *x dan* y
x = data pusat klaster
y = data pada atribut
i = setiap data
n = jumlah data,
x_i = data pada pusat klaster ke *i*
y_i = data pada setiap data ke *i*

C. *Minkowski Distance*

Minkowski distance merupakan sebuah metrik dalam ruang vektor di mana suatu norma didefinisikan (*normed vector space*) sekaligus dianggap sebagai generalisasi dari *Euclidean distance* dan *Manhattan distance*. Dalam pengukuran jarak objek menggunakan *minkowski distance* biasanya digunakan nilai *p* adalah 1 atau 2. Berikut rumus yang digunakan menghitung jarak dalam metode ini.

$$d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} \quad (3)$$

dimana,
d = jarak antara *x dan* y
x = data pusat klaster
y = data pada atribut
i = setiap data
n = jumlah data,
x_i = data pada pusat klaster ke *i*
y_i = data pada setiap data ke *i*
p = power

D. *Teknik Analisis Cluster*

Proses pengelompokan (*clustering*) data dilakukan melalui tahapan umum dari algoritma *K-Means Clustering*. Sebelum proses clustering dimulai, terlebih dahulu dilakukan normalisasi data dengan normalisasi *Min-Max*. *Min-Max* merupakan metode normalisasi dengan melakukan transformasi *linier* terhadap data asli. Proses normalisasi bertujuan untuk memetakan nilai dari masing-masing variabel ke dalam rentang yang sama yakni rentang [0,1], sehingga pada saat proses perhitungan nilai *similarity*, masing-masing variabel memberikan tingkat kepentingan yang sama (memberikan pengaruh yang sama. Normalisasi ini dapat dilakukan dengan menggunakan persamaan berikut:

$$x' = \frac{x - nilai_{min}}{nilai_{max} - nilai_{min}} \quad (4)$$

dimana,
x = data per kolom
$nilai_{min}$ = nilai minimum dari data per kolom
$nilai_{max}$ = nilai maksimum dari data perkolom

Setelah itu proses dilanjutkan dengan menentukan jumlah *cluster K* menggunakan pendekatan *rule-of-thumb* dengan persamaan:

$$k = \sqrt{\frac{n}{2}}. \quad (5)$$

dimana,
n = jumlah objek yang akan di kelompokkan, dan
k = jumlah *cluster*.

Tahapan pertama, menentukan nilai *centroid* awal dengan menggunakan rumus berikut:

$$C_i = \frac{\sum_{i=1}^{n} x_i \in s_i}{n} \quad (6)$$

dimana,
C_i = *centroid* baru ke *i*
s_i = objek ke *i*
x_i = nilai pada objek ke *i*
n = jumlah data pada tiap kelompok

Tahapan kedua, menghitung jarak objek dengan *centroid* dengan menggunakan beberapa metode. Pada penelitian ini digunakan metode *Euclidean Distance*, *Manhattan Distance*, dan *Minkowski Distance*. Setelah jarak dihitung, selanjutnya dilakukan uji homogenitas klaster. Pengujian homogenitas klaster dapat ditentukan berdasarkan nilai koefisien *silhouette* yang dapat diperoleh melalui beberapa tahapan meliputi: perhitungan rata-rata jarak dari suatu objek, misalkan *i* dengan semua objek lain yang berada dalam satu klaster dengan menggunakan persamaan di bawah ini.

$$a_i = \frac{1}{|A| - 1} \sum_{j \in A, i \neq j} d(i,j) \quad (7)$$

dimana:
$|A|$ = banyaknya data dalam klaster *A*
i, j = indeks dari dokumen
$d(i,j)$ = jarak antara dokumen ke *i* dengan dokumen ke *j*.

Setelah rata-rata jarak dihitung, selanjutnya dihitung rata-rata jarak dari dokumen *i* tersebut dengan semua dokumen di klaster lain, dan diambil nilai terkecilnya dengan menggunakan rumus berikut.

$$\boldsymbol{d(i,C) = \frac{1}{|A|} \sum_{j \in C} d(i,j)}$$

dimana, $d(i,C)$ adalah jarak rata-rata objek i dengan semua objek pada klaster lain C dimana $A \neq C$.

$$b(i) = \min_{C \neq A} d(i,C) \quad (8)(9)$$

Setelah nilai terkecil didapatkan, selanjutnya dihitung nilai *silhouette coefficient*-nya dengan persamaan berikut.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (10)$$

Ketiga, dilakukan pengelompokkan objek dan pengujian konvergensi antara kelompok data baru dan kelompok data pada proses sebelumnya, jika sama (*convergen*), maka proses *clustering* selesai. Jika tidak, maka harus dilakukan iterasi dimulai dari penentuan pusat klaster baru. Setelah klaster dihasilkan, selanjutnya dilakukan pengujian menggunakan *Chi-Square* untuk mengetahui perbandingan antara frekuensi observasi dengan frekuensi harapan yang didasarkan pada hipotesis pada data yang diambil untuk diamati. Pada penelitian ini, nilai interpretasi dari *Chi-Square* akan didasarkan pada perbandingan antara nilai-nilai aritmatika *Chi-Square* dengan nilai tabel *Chi-Square*. Nilai yang lebih besar dari aritmatika *Chi-Square* berarti ada perbedaan yang signifikan antara frekuensi observasi dengan frekuensi harapan, begitu juga sebaliknya [12]. Persamaan yang dapat digunakan adalah seperti berikut:

$$X^2 = \sum \frac{(F_o - F_h)^2}{F_h} \quad (11)$$

dimana, *X2* = Nilai *Chi-square*, *Fo* = Nilai yang diobservasi, dan *Fh* = Nilai yang diharapkan.

E. Dataset

Dataset yang digunakan untuk pengujian merupakan data jumlah siswa dan ketersediaan Guru di wilayah Kota Tegal yang dapat diperolah dari dapodik (data pokok pendidikan) wilayah Jawa Tengah. Selain itu, dalam penelitian ini juga menggunakan data rasio kebutuhan Guru sesuai dengan angka ideal rasio siswa-guru yang ada dalam Peraturan Pemerintah Nomor 74 tahun 2008 tentang Guru.

IV. HASIL DAN PEMBAHASAN

Proses klastering menghasilkan 9 (sembilan) klaster dengan jumlah anggota yang berbeda, tergantung dengan metode pengukuran jarak yang digunakan, seperti ditunjukkan pada Tabel 1. Berdasarkan hasil uji *Chi-Square* yang telah dilakukan (dapat dilihat pada Tabel II, Tabel III, dan Tabel IV), ditemukan bahwa dari ketiga metode yang digunakan sebagai pengukur jarak memiliki tingkat akurasi yang hampir sama, seperti ditunjukkan pada Tabel V.

Pada Tabel II, Tabel III, dan Tabel IV memiliki atribut dimana K = klaster, F_o (A): Nilai ketersediaan Guru, F_o (B): Nilai kebutuhan Guru, F_h: Nilai yang harapkan, A^2: Chi-square ketersediaan Guru, B^2: Chi-square kebutuhan Guru, X^2: Nilai Chi-square, L: Label disparitas (W=WAJAR, T=TINGGI).

TABEL I
HASIL *CLUSTERING*

Klaster	Jumlah Sekolah		
	Euclidean Distance	**Manhattan Distance**	**Minkowski Distance**
1	6	13	6
2	14	5	13
3	17	15	13
4	41	47	47
5	23	14	15
6	14	15	15
7	7	8	7
8	9	8	9
9	30	36	36
Total	161	161	161

TABEL II
HASIL PENGUJIAN CHI-SQUARE (MINKOWSKI DISTANCE)

K	F_o (A)	F_o (B)	F_h	A^2	B^2	X^2	L
1	52	15	33.5	10.216	10.216	20.433	T
2	106	70	88	3.6818	3.6818	7.3636	T
3	350	319	334.5	0.7182	0.7182	1.4365	W
4	403	428	415.5	0.3761	0.3761	0.7521	W
5	549	491	520	1.6173	1.6173	3.2346	W
6	286	245	265.5	1.5829	1.5829	3.1657	W
7	92	47	69.5	7.2842	7.2842	14.568	T
8	565	543	554	0.2184	0.2184	0.4368	W
9	300	269	284.5	0.8445	0.8445	1.6889	W

TABEL III
HASIL PENGUJIAN CHI-SQUARE (MANHATTAN DISTANCE)

K	F_o (A)	F_o (B)	F_h	A^2	B^2	X^2	L
1	101	67	84	3.4405	3.4405	6.881	T
2	47	12	29.5	10.381	10.381	20.763	T
3	408	378	393	0.5725	0.5725	1.145	W
4	403	428	415.5	0.3761	0.3761	0.7521	W
5	535	475	505	1.7822	1.7822	3.5644	T
6	286	245	265.5	1.5829	1.5829	3.1657	W
7	102	53	77.5	7.7452	7.7452	15.49	T
8	521	500	510.5	0.216	0.216	0.4319	W
9	300	269	284.5	0.8445	0.8445	1.6889	W

TABEL IV
HASIL PENGUJIAN CHI-SQUARE (EUCLIDEAN DISTANCE)

K	F_o (A)	F_o (B)	F_h	A^2	B^2	X^2	L
1	52	15	33.5	10.216	10.216	20.433	T
2	118	74	96	5.0417	5.0417	10.083	T
3	366	324	345	1.2783	1.2783	2.5565	W
4	327	358	342.5	0.7015	0.7015	1.4029	W
5	767	696	731.5	1.7228	1.7228	3.4457	W
6	148	144	146	0.0274	0.0274	0.0548	W
7	102	52	77	8.1169	8.1169	16.234	T
8	565	543	554	0.2184	0.2184	0.4368	W
9	258	221	239.5	1.429	1.429	2.858	W

TABEL V
TINGKAT AKURASI METODE

No	Distance Measure	Coefficient silhouette (Rata-rata)	Akurasi
1	Euclidean	0.514	84.47%
2	Manhattan	0.550	83.85%
3	Minkowski	0.560	83.85%

Pada penelitian ini, *dataset* yang digunakan berjumlah 161 data saatuan pendidikan tingkat dasar, menengah, dan tingkat atas di Kota Tegal. Data tersebut diambil dari data pokok pendidikan (dapodik) wilayah Jawa Tengah. Sedangkan penentuan jumlah klaster *K* ditentukan menggunakan pendekatan *rule-of-thumbs*.

$$K = \sqrt{\frac{n}{2}}.$$

Sehingga:

$K = \sqrt{\frac{161}{2}} = 8.97$ atau bisa dibulatkan menjadi *K*=9.

Dari hasil uji homogenitas klaster (untuk mengetahui struktur klaster) yang telah dilakukan menggunakan *silhoutte coefficient*, diperoleh hasil bahwa nilai rata-rata *silhoutte* adalah 0.514 (menggunakan *euclidean*), 0.550 (menggunakan *manhattan*), dan 0.560 (menggunakan *minkowski*). Hal ini menunjukkan bahwa klastering dengan menggunakan ketiga metode pengukuran jarak tersebut memiliki struktur medium (baik).

Pengujian akurasi dari ketiga metode pengukuran jarak tersebut diawali dengan memberikan label disparitas (ketersediaan dengan kebutuhan) pada setiap klaster dan anggota dalam klaster menggunakan *chi-square*. Karena dalam hal ini digunakan 2 (dua) kategori, yaitu ketersediaan dan kebutuhan, maka nilai derajat kebebasan adalah *(2-1) = 1*. Dengan demikian, nilai derajat kebebasan adalah 1 dan toleransi kesalahan 0.5, maka nilai *Chi-square* adalah 3.841, oleh karena itu berlaku aturan dalam pemberian label yaitu: hasil pelabelan dianggap wajar jika nilai *chi-square* kurang dari 3.841, dan selain itu hasil pelabelan dianggap tinggi. Selanjutnya dilakukan pengujian tingkat akurasi dengan cara membagi jumlah label terprediksi benar dengan jumlah keseluruhan label. Hasil evaluasi menunjukkan bahwa pelabelan memiliki tingkat akurasi yang tinggi, yaitu 84.47% (*euclidean*), 83.85% (*manhattan*), dan 83.85% (*minkowski*).

V. KESIMPULAN

Perbandingan akurasi metode pengukuran jarak (*euclidean*, *manhattan*, dan *minkowski*) untuk pelabelan klaster status disparitas kebutuhan Guu telah dilakukan dan memberikan nilai atau tingkat akurasi yang tinggi, yaitu 84.47% (untuk metode *euclidean distance*), 83.85% (untuk metode *manhattan distance*), dan 83.85% (untuk metode *minkowski*). Dengan demikian, dapat disimpulkan bahwa metode *euclidean* meupakan metode terbaik untuk diterapkan dalam algoritma *K-Means Clustering*. Selanjutnya, klaster label klaster dapat digunakan mengidentifikasi status disparitas Guru untuk masing-masing sekolah di Kota Tegal. Berdasarkan pelabelan pada klaster tesebut, maka dapat diketahui bahwa sekolah dengan kondisi ketersediaan Guru yang masih sangat kurang (kategori label disparitas TINGGI) dan perlu mendapatkan perhatian lebih adalah SMP Atmaja Wacana, SMKN 3 Tegal, SMAS Muhammadiyah, SMAS Pancasakti Tegal, SMKS Muhammadiyah 1 Kota Tegal, dan SMP IC Bias Assalam.

DAFTAR PUSTAKA

[1] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining (2nd Edition)*, 2nd ed. New York: Pearson, 2018.

[2] M. Nishom, "Implementasi Metode K-Means berbasis Chi-Square pada Sistem Pendukung Keputusan untuk Identifikasi Disparitas Kebutuhan Guru," *J. Sist. Inf. Bisnis*, vol. 8, no. 2, pp. 1–8, 2018.

[3] S. Saraswathi and M. I. Sheela, "A Comparative Study of Various Clustering Algorithms in Data Mining," vol. 3, no. 11, pp. 422–428, 2014.

[4] R. Awasthi, A. K. Tiwari, and S. Pathak, "Empirical Evaluation On K Means Clustering With Effect Of Distance Functions For Bank Dataset," *Int. J. Innov. Technol. Res.*, vol. 1, no. 3, pp. 233–235, 2013.

[5] A. Singh, A. Rana, and A. Yadav, "K-means with Three different Distance Metrics," *Int. J. Comput. Appl.*, vol. 67, no. 10, pp. 13–17, 2013.

[6] K. Kouser, "A comparative study of K Means Algorithm by Different Distance Measures," *Int. J. Innov. Res. Comput.*, vol. 1, no. 9, pp. 2443–2447, 2013.

[7] D. Sinwar and R. Kaushik, "Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 2, no. 5, pp. 270–274, 2014.

[8] D. J. Bora and A. K. Gupta, "Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab," *Eff. Differ. Distance Meas. Perform. K-Means Algorithm An Exp. Study Matlab*, vol. 5, no. 2, pp. 2501–2506, 2014.

[9] H. Prasetyo and A. Purwariati, "Comparison of Distance Measures for Clustering Data with Mix Attribute Types," in *International Conference on Information Technology Systems and Innovation*, 2014.

[10] A. Singh, J. Agarwal, and A. Rana, "Performance Measure of Similis and FPGrowth Algo rithm," *Int. J. Comput. Appl.*, vol. 62, no. 6, pp. 25–31, 2013.

[11] H. Anton, *Elementary Linear Algebra*, 7th ed. New Jersey: Wiley, 1993.

[12] R. Stine, *Statistics for Business Decision Making and Analysis with Chi-Square Tests*. New York: Pearson, 2011.