*Jurnal Ilmiah Matematika dan Terapan*
Volume 20 Nomor 2 Desember 2023 (Halaman 147 - 153)
doi: https://doi.org/10.22487/2540766X.2023.v20.i2.16871

JIMT
Jurnal Ilmiah Matematika dan Terapan
ISSN : 2540 – 766X

# ANALISIS KOMPARASI ALGORITMA CLUSTERING BERBASIS PARTISI UNTUK DATA NUMERIK DAN DATA KATEGORIKAL

**Desy Lusiyanti[1], Imam Al Fajri[2], Andri[3], Mohammad Fajri[4]**

[1,2,3,]Program Studi Matematika, Jurusan Matematika, FMIPA, Universitas Tadulako
[4,]Program Studi Statistika, Jurusan Matematika, FMIPA, Universitas Tadulako
[1]desy.lusiyanti@untad.com, [2]iman.alfajri@gmail.com, [3]andri90@untad.ac.id, [4]m.fajri@untad.ac.id

**ABSTRACT**

In practice, not always all data features are of the numeric type or categorical type. Differences in features in data become a problem in determining the method to be used. One way that is often used to overcome this problem is to change one of the feature values by adjusting the method that will be used. For example, in cluster analysis, several algorithms are often used, including K-Means and K-Modes. These two methods have differences in the features used. K-Means uses a numeric data type while K-Modes uses a categorical data type. In this study, a comparison was carried out between the K-Means and K-Modes methods to cluster heart disease patients. The dataset used in this research is medical record data from heart disease patients at Undata Palu Hospital. The research results show that the two methods compared have a good level of accuracy, namely 84.47% (for the K-Means method), and 83.85% (for the K-Modes method).

**Keywords : Clustering , K-Means, K-Modes.**

**ABSTRAK**

Dalam prakteknya, tidak selalu semua fitur data bertipe numerik ataupun bertipe kategorik. Perbedaan fitur pada suatu data menjadi permasalahan dalam menentukan metode yang akan digunakan. Salah satu cara yang sering digunakan untuk mengatasi permasalahan tersebut yaitu mengubah salah-satu dari nilai fitur dengan menyesuaikan metode yang akan digunakan. Misalkan dalam analisis cluster, terdapat beberapa algoritma yang sering digunakan diantaranya adalah K-Means dan K-Modes. Kedua metode ini memiliki perbedaan dari fitur yang digunakan. K-Means menggunakan tipe data numerik sedangkan K-Modes menggunakan tipe data kategorik. Dalam penelitian ini dilakukan komprasi antara metode K-Means dan K-Modes untuk mengclusterkan pasien penyakit jantung. Dataset yang digunakan dalam penelitian ini adalah data rekam medis pasien penyakit jantung RSUD Undata palu. Hasil penelitian menunjukkan bahwa dari kedua metode yang dibandingkaan memiliki tingkat akurasi yang baik, yaitu 84.47% (untuk metode K-Means), dan 83.85% (untuk metode K-Modes).

**Kata kunci : Klastering, K-Means, K-Modes.**

## I. PENDAHULUAN

Dewasa ini, kecermatan dan ketepatan metode merupakan hal yang sangat penting untuk diperhatikan. Contohnya untuk kasus clustering, hasil clustering data dengan beberapa metode yang telah dikembangkan akan mempengaruhi sejumlah keputusan. Data Mining berpotensi untuk membantu mengambil sebuah keputusan yang sederhana namun akurat (Fauset, 1993). Berbagai metode clustering dikenal dalam data mining, seperti K-Means, K-Modes, Fuzzy K-Means dan lain sebagainya (C Raju, 2018).

Analisis cluster merupakan salah satu teknik data mining yang bertujuan untuk mengidentifikasi sekelompok obyek yang mempunyai kemiripan karakteristik tertentu yang dapat dipisahkan dengan kelompok obyek lainnya, sehingga obyek yang berada dalam kelompok yang sama relatif lebih homogen daripada obyek yang berada pada kelompok yang berbeda (Irawan, 2008). Jumlah kelompok yang dapat diidentifikasi tergantung pada banyak dan variasi data obyek. Tujuan dari pengelompokan sekumpulan data obyek kedalam beberapa kelompok yang mempunyai karakteristik tertentu dan dapat dibedakan satu sama lainnya adalah untuk analis dan interprestasi lebih lanjut sesuai dengan tujuan penelitian yang dilakukan.

Model yang diambil diasumsikan bahwa data yang dapat digunakan adalah data yang berupa interval, frekuensi dan biner. Set data obyek mempunyai peubah dengan tipe yang sejenis yang tidak bercampur antara tipe yang satu dengan lainnya. Dalam artian data kategorik tidak bercampur dengan data numerik. Hal ini dapat memberikan akurasi yang buruk terhadap metode. Pemakaian metode dan tipe data haruslah sesuai agar performa dari metode tersebut menghasilkan nilai yang baik.

Dengan adanya beberapa metode yang tersedia, permasalahan yang sering muncul adalah jenis metode yang harus dipilih. Setiap metode tentunya memiliki kelebihan dan kekurangan. Dalam beberapa kasus, kita dapat menerapkan beberapa teknik atau satu teknik dengan parameter yang berbeda selanjutnya model tersebut dibandingkan dengan melihat tingkat error yang dihasilkan. Error yang kecil mempunyai akurasi tertinggi. Namun satu hal yang perlu ditekankan bahwa tidak ada model terbaik untuk semua kasus atau data yang ada. Oleh karena itu, perbandingan antara metode yang satu dengan yang lain merupakan hal yang menarik untuk diteliti.

Penelitian ini akan melakukan perbandingan algoritma clustering yaitu K-Means dan K-Modes dengan menggunakan uji parametrik dengan t-test agar dapat menghasilkan perbandingan metode yang lebih baik untuk data set kategorik dan numerik. Data yang digunakan adalah data rekam medis pasien jantung yang diperoleh dari data sekunder RS Undata Palu.

## II. METODE PENELITIAN

Data yang digunakan pada penelitian ini adalah data rekam medis pasien penyakit jantung koroner yang diperoleh dari RSUD Undata Palu. Metode yang digunakan dalam pengambilan data sampel ini adalah selain mengambil data yang sudah tersedia di data administrasi pada rumah sakit tersebut juga dilakukan wawancara terhadap dokter spesialis penyakit jantung untuk mengetahui gambaran secara umum penyakit jantung. Data pasien penyakit jatung diperoleh sebanyak 50 orang pasein yang terdiri dari sembilan faktor resiko yang diperoleh berdasarkan catatan rekam medik pasien seperti pada tabel dibawah ini Data-data yang digunakan ditampilkan pada Tabel 1.

Tabel 1 : Faktor Resiko Penyakit Jantung Koroner

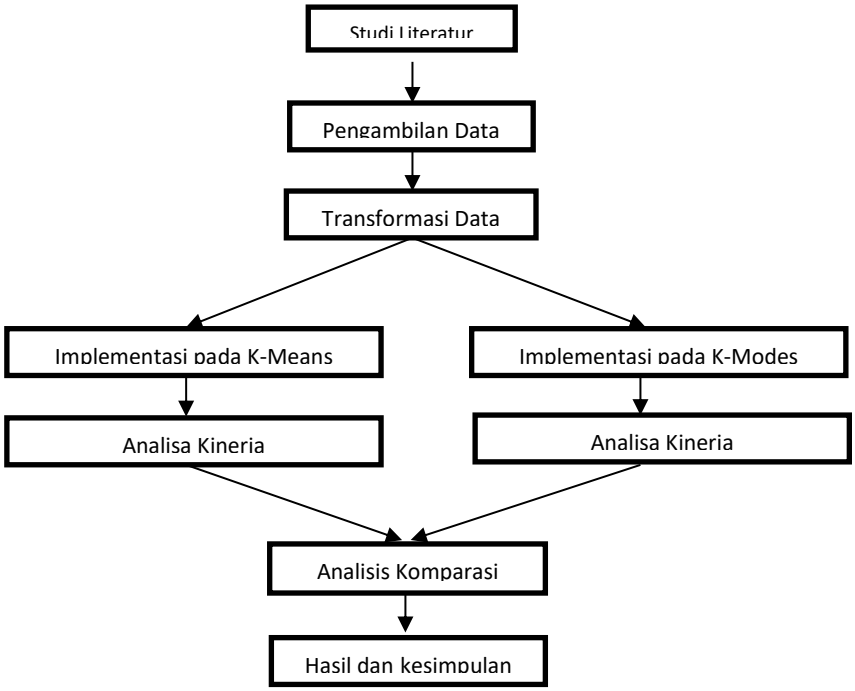
| Atribut | Keterangan | Class |
|---|---|---|
| $a_1$ | Jenis Kelamin | Kategorik |
| $a_2$ | Usia | Numerik |
| $a_3$ | Pekerjaan | Kategorik |
| $a_4$ | Glukosa | Numerik |
| $a_5$ | Kolesterol Total | Numerik |
| $a_6$ | Trigliserida | Numerik |
| $a_7$ | Lactic Dhydrogenas | Numerik |
| $a_8$ | Low Density Lipoprotein | Numerik |
| $a_9$ | Asam Urat | Numerik |

Dari 9 atribut dataset tersebut, data dengan bentuk kategorik akan ditransformasi menjadi bentuk numerik untuk diolah dengan menggunakan metode K-means dan data dengan bentuk numerik akan ditransformasi menjadi bentuk kategorik untuk diolah dengan menggunakan metode k-Modes. Berikut aturan transformasi data yang digunakan

| | | | |
|---|---|---|---|
| $a_1$ = Jenis Kelamin | 1 | = Laki-laki | |
| | 2 | = Perempuan | |
| $a_2$ = Usia | Pertengahan | = 45-59 | |
| | Lanjut Usia | = 60-74 | |
| | Tua | = 75-90 | |
| | Sangat Tua | > 90 | *(Sumber: Who )* |
| $a_3$ = Pekerjaan | 0 | = PNS | |
| | 1 | = Wiraswasta | |
| | 2 | = URT/IRT | |
| | 3 | = Pensiunan | |
| | 4 | = Lain-lain | |
| $a_4$ = Kadar Glukosa | Normal = 80-139 | | |
| | Sedang = 140-199 | | |

| | | |
|---|---|---|
| | Buruk | > 200 |
| | | (*sumber :* Dinas Kesehatan) |
| $a_5$ = Kadar Kolesterol | Normal | < 200 |
| | Ambang Batas Tinggi = 200-239 | |
| | Tinggi | > 240 (*sumber :* Dinas Kesehatan) |
| $a_6$ = Kadar Trigliserida | Normal | < 200 |
| | Tinggi | = 200-400 |
| | Sangat Tinggi | > 400 (*sumber :* Dinas Kesehatan) |
| $a_7$ = Kadar Lactic Dhydrogenas : | Normal | > 60 |
| | Perbatasan | = 40-59 |
| | Bahaya | < 40 (*sumber :* Dinas Kesehatan) |
| $a_8$ = kadar Density Lipoprotein : | Normal | < 130 |
| | Perbatasan | = 130-159 |
| | Bahaya | > 160 (*sumber :* Dinas Kesehatan) |
| $a_9$ = Asam Urat | Laki-Laki | = 3,5 – 7 (normal) |
| | Perempuan | = 2,6-6 (normal) |
| | | (*sumber :* Dinas Kesehatan) |

Secara umum metodelogi penelitian dapat dilihat pada Gambar 1.



## III. HASIL DAN PEMBAHASAN

Pengujian dilakukan dalam dua cara. Pada cara pertama, dilakukan pengelompokkan data pasien menggunakan algoritma pengelompokkan K-Means dan pada cara kedua, K-Modes diimplementasikan pada algoritma pengelompokkan pasien terdiagnosis penyakit jantung. Jumlah cluster untuk kedua metode sama yaitu 2 cluster. Distance metrik yang digunakan adalah Euclidean Distance dan metrik pencocokan.

### 3.1. Metode K-Means

Berikut langkah-langkah dalam pengerjaan dengan menggunakan metode K-Means :

1. Inisialisasi

   Menentukan banyaknya klaster. klaster yang akan dibuat (k), nilai k = 2
2. Memilih 2 objek secara acak dan terpilih objek ke-13, dan ke-67 sebagai centroid iterasi pertama yang disajikan pada tabel 5.3 berikut

Tabel 2 : Centroid Awal Klaster

| Centroid (C ) | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 74 | 0 | 147 | 109 | 59 | 25 | 72 | 10.3 |
| 2 | 1 | 51 | 4 | 360 | 274 | 289 | 49 | 167 | 2.6 |

3. Menghitung jarak setiap data ke centroid terdekat. Centroid terdekat akan menjadi cluster yang diikuti oleh data tersebut. Berikut contoh perhitungan jarak ke setiap centroid pada data ke-1 :

$$d(x_1, c_1) = \sqrt{\sum_{i=1}^{r} (x_{1i} - c_{1i})^2}$$

$$= \sqrt{(2-1)^2 + (65-74)^2 + (4-0)^2 + (120-147)^2 + (213-109)^2 + (126-59)^2 + } \\ = \sqrt{+(42-25)^2 + (146-72)^2 + (4.5-10.3)^2}$$

$$d(x_1, c_1) = 148.09$$

$$d(x_1, c_1) = \sqrt{\sum_{i=1}^{r} (x_{1i} - c_{1i})^2}$$

$$= \sqrt{(1-1)^2 + (65-51)^2 + (4-4)^2 + (120-360)^2 + (213-274)^2 + (126-289)^2 + } \\ = \sqrt{+(42-49)^2 + (146-167)^2 + (4.5-2.6)^2}$$

$$d(x_1, c_1) = 297.62$$

4. Menentukan anggota dari masing-masing data berdasarkan jarak terdekat terhadap centroid klaster (C) dengan rumus jarak terdekat yaitu min $\{d(x_1, c_1), d(x_1, c_2)\}$.
5. Memperbarui centroid masing-masing klaster berdasarkan mean (nilai rata-data) dari setiap variabel.

6. Menghitung kembali jarak setiap objek terhadap centroid baru seperti pada langkah (3). Kemudian menentukan kembali anggota dari masing-masing klaster berdasarkan jarak terdekat terhadap centroid klaster terbaru seperti pada langkah (4).
7. Mengulangi langkah 3 dan 4 hingga kondisi konvergen tercapai, yaitu (a) perubahan fungsi objektif sudah di bawah ambang batas yang diinginkan; atau (b) tidak ada data yang berpindah cluster; atau (c) perubahan posisi centroid sudah di bawah ambang batas yang ditetapkan (S Nagari, 2020).

### 3.2. Metode K-Modes

Berikut ini disajikan langkah-langkah perhitungan secara manual proses pengklasteran dengan metode k-Modes. Proses pengklasteran dilakukan dengan inisialisasi nilai k yaitu 2. Hal ini dikarenakan data yang diperoleh merupakan data yang sudah diketahui kelasnya 2 (Penyakit jantung dan penyakit jantung koroner). Berikut adalah langkah-langkah pengerjaannya:

1. Tentukan banyaknya klaster yang akan dibuat (k), nilai k = 2
2. Memilih 2 objek secara acak dan terpilih objek ke-13 , dan ke-67 sebagai centroid iterasi pertama yang disajikan pada Tabel 4.

   Tabel 4 : Centroid Awal Klaster

| Centroid (C ) | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | $a_9$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Laki Laki | Lanjut Usia | Pensiunan | Sedang | Normal | Normal | Bahaya | Normal | Tidak Normal |
| 2 | Perempuan | Pertengahan | Peg. Swasta | Buruk | Tinggi | Tinggi | Perbatasan | Bahaya | Normal |

3. Hitung jarak dari masing-masing objek dengan centroid iterasi pertama menggunakan ukuran ketidaksamaan pencocokan $d(X,Y) = \sum_{j=1}^{r} \in (x_j, y_j)$ dengan $\in$ ( ) adalah nilai pencocokan dengan persamaan $\in (x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$. Sebagai contoh jarak antara data ke-1 terhadap centroid tiap klaster yaitu $C_1$ dan $C_2$ adalah sebagai berikut:

$$d(x_1, c_1) = \in (x_{11}, c_{11}) + \in (x_{12}, c_{12}) + \in (x_{13}, c_{13}) + \in (x_{14}, c_{14}) + (x_{15}, c_{15}) + \in (x_{16}, c_{16}) + \in (x_{17}, c_{17}) + \in (x_{18}, c_{18}) + \in (x_{19}, c_{19})$$

$$= \in (perempuan, laki - laki) + \in (lanjut\ usia, lanjut\ usia) + \in (urt, pensiunan\ abri) + \in (normal, sedang) + \in (ambang\ batas\ tinggi, normal) + \in (normal, normal) + \in (perbatasan, bahaya) + \in (perbatasan, normal) + \in (normal, tidak\ normal)$$

$$= 1 + 0 + 1 + 1 + 1 + 0 + 1 + 1 + 1 = 7$$

$$d(x_1, c_2) = \in (x_{11}, c_{21}) + \in (x_{12}, c_{22}) + \in (x_{13}, c_{23}) + \in (x_{14}, c_{24}) + (x_{15}, c_{25}) + \in (x_{16}, c_{26}) + \in (x_{17}, c_{27}) + \in (x_{18}, c_{28}) + \in (x_{19}, c_{29})$$

$$= \in (perempuan, perempuan) + \in (lanjut\ usia, pertengahan) + \in (urt, peg\ swasta) + \in (normal, buruk) + \in (ambang\ batas\ tinggi, tinggi) + \in (normal, tinggi) + \in (perbatasan, perbatasan) + \in (perbatasan, bahaya) + \in (normal, normal)$$

$$= 0 + 1 + 1 + 1 + 1 + 1 + 0 + 1 + 0 = 6$$

4. Menentukan anggota dari masing-masing data berdasarkan jarak terdekat terhadap centroid klaster (C) dengan rumus jarak terdekat yaitu min $\{d(x_1, c_1), d(x_1, c_2)\}$.
5. Memperbarui centroid masing-masing klaster berdasarkan modus dari setiap variabel.
6. Hitung ulang jarak setiap objek terhadap centroid baru seperti pada langkah (3). Kemudian tentukan ulang anggota dari masing-masing klaster berdasarkan jarak terdekat terhadap centroid klaster terbaru seperti pada langkah (4).

### 3.3. Perbandingan Hasil Clustering K-Means dan K-Modes

Dari pengujian yang telah dilakukan untuk metode K-Means dan K-Modes dalam mengelompokkan data pasien terdiagnosis penyakit jantung, didapat iterasi clusternya. Perbandingan iterasi, dapat dilihat pada Tabel 5.

Tabel 5 : Hasil Perbandingan Iterasi K-Means dan K-Modes

| | K-Means | K-Modes |
|---|---|---|
| Iterasi | 5 | 3 |
| Accuracy | 84.47% | 83.35% |

## IV. KESIMPULAN

Dari hasil penelitian di atas, diperoleh hasil clustering dengan menggunakan metode k-means dan dengan menggunakan metode k-modes. Hasil penelitian menunjukkan bahwa dari kedua metode yang dibandingkan memiliki tingkat akurasi yang baik, yaitu 84.47% untuk metode K-Means, dan 83.85% untuk metode K-Modes.

## DAFTAR PUSTAKA


[1]. C. Raju, E. Philipsy, S. Chacko, L. Padma Suresh, dan S. Deepa Rajan. (2018). “A Survey on Predicting Heart Disease Using Data Mining Techniques,” Proc. IEEE Conf. Emerg. Devices Smart Syst. ICEDSS 2018, No. March, Hal. 253–255.

[2]. Fausett, L.V. (1993). Fumdamental of Neural Network: Architectures, Algorithm, and Application. Prentice Hall, 1st edition. ISBN-13: 978-0133341867.

[3]. Irawan, M. I. (2008). Exploratory Data Analysis dengan JST-Kohonen SOM: Struktur Tingkat Kesejahteraan Darah Tk II se-Jawa Timur. ITS, Surabaya.

[4]. S. S. Nagari dan L. Inayati. (2020) . “Implementation of Clustering Using K-Means Method to Determine Nutritional Status,” J. Biometrika dan Kependud., Vol. 9, No. 1, Hal. 62–68.