Teknik Informatika
Fakultas Teknik
Universitas Muhammadiyah Jakarta
Jl. Cempaka Putih Tengah 27
Jakarta Pusat, Indonesia 10510
T. 62 21 4256024, 4244016 ext 207
E. jurnal.justit@ftumj.ac.id
H. https:
https://jurnal.umj.ac.id/index.php/just-it

DIAGNOSA PENYAKIT KANKER PAYUDARA MENGGUNAKAN METODE K-MEANS CLUSTERING

Emi Susilowati[1], Amelia Tri Hapsari[2], Muhammad Efendi[3], dan Priadhana Edi Kresnha[4]

[1,2,3,4]Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jakarta

emi.susilowati@ftumj.ac.id

Abstrak

Kanker payudara merupakan kondisi kanker yang muncul di daerah payudara. Kanker jenis ini sering dialami oleh wanita dengan ciri khas dari kanker payudara yaitu munculnya benjolan yang tidak biasa di area payudara. Salah satu metode pemeriksaan kanker payudara adalah mammografi yang merupakan metode skrining yang akan mengidentifikasi kanker payudara berdasarkan gejala-gejala yang muncul. Pada penelitian ini akan dilakukan pengklasteran diagnosa keadaan pasien kanker payudara benign (jinak) dan malign (ganas) berdasarkan faktor-faktor yang mempengaruhi kanker payudara dengan menggunakan metode K-Means Clustering. Hasil dari clustering tersebut akan menentukan pada tiap-tiap pasien apakah pasien tersebut tergolong kanker payudara jinak (benign) ataupun kanker payudara ganas (malignant). Proses clustering mengambil data 41 pasien dengan 10 gejala yang mempengaruhi kanker payudara. Berdasarkan perhitungan menggunakan metode k-means clustering menghasilkan dua cluster dimana cluster 1 (C1) terdapat 11 data yaitu data ke-2, 4, 7, 9, 10, 14, 15, 17, 19, 27, dan 40 yang tergolong dalam penyakit kanker payudara ganas. Cluster 2 (C2) terdapat 30 data yaitu data ke-1, 3, 5, 6, 8, 11, 12, 13, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, dan 41 yang tergolong dalam penyakit kanker payudara jinak.
Kata Kunci: *Kanker Payudara, Mammografi, Clustering, K-Means*

Abstract

Breast cancer is a condition of cancer that occurs in the breast area. This type of cancer is often experienced by women with a distinctive feature of breast cancer which is the appearance of an unusual lump in the breast area. One of the breast cancer screening methods is mammography, which is a screening method that will identify breast cancer based on emerging symptoms. In this study, the diagnosis of benign and malignant breast cancer patients is based on factors that influence breast cancer using the K-Means Clustering method. The results of the clustering will determine in each patient whether the patient is benign or benign or malignant breast cancer. The clustering process took data on 41 patients with 10 symptoms affecting breast cancer. Based on calculations using the k-means clustering method produces two clusters where cluster 1 (C1) contains 11 data, namely 2, 4, 7, 9, 10, 14, 15, 17, 19, 27, and 40 that belong to the disease

malignant breast cancer. Cluster 2 (C2) contains 30 data, 1, 3, 5, 6, 8, 11, 12, 13, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, and 41 that belong to benign breast cancer.
Keywords: *Breast Cancer, Mammography, Clustering, K-Means*

PENDAHULUAN

Kanker payudara adalah suatu kanker dimana bertumbuhnya serta berkembangnya sebuah sel-sel jaringan yang mengerikan yang tumbuh di area payudara. Kanker payudara merupakan kanker yang sering terjadi kedua di dunia dan kanker yang paling sering dirasakan oleh diantara wanita dengan perkiraan 1,67 juta kasus kanker baru yang didiagnosis pada tahun 2012 (25% dari semua kanker). Penderita kanker payudara banyak dialami oleh wanita yang tinggal di daerah kurang berkembang, dibandingkan yang tinggal di daerah maju (InfoDATIN, 2016).

Menurut data dari GLOBOCAN (IARC) tahun 2012 bahwa kanker payudara merupakan penyakit kanker dengan presentase kasus baru tertinggi dan presentase penyebab kematian tertinggi. Data GLOBOCAN menunjukan bahwa kanker payudara memiliki presentase kematian penderita kanker payudara yang jauh lebih rendah dibandingkan dengan kasus baru, sehingga jika penyakit kanker payudara dapat dicegah sejak dini maka kemungkinan sembuh pada kanker payudara juga akan lebih tinggi (InfoDATIN, 2016).

Salah satu metode pemeriksaan kanker payudara yaitu metode mammografi. Metode mammografi merupakan metode skrining yang dapat mengidentifikasi kanker berdasarkan gejala-gejala fisik penyakit tersebut muncul. Namun, terkadang hasil mammografi dianggap tidak meyakinkan. Oleh karena itu diperlukan perangkat tambahan yang dapat meyakinkan pengidentifikasian kanker payudara apakah hasil mammografi seorang pasien termasuk ke kelas jinak, yang memiliki harapan kecil untuk terkena kanker payudara atau kelas ganas yang di nyatakan kanker payudara yang parah (Keleş, Keleş, & Yavuz, 2011).

Penelitian terhadap diagnosa penyakit kanker payudara dengan menggunakan berbagai macam metode telah banyak dilakukan. Pada penelitian yang diakukan oleh (Aliady, Tuasikal, & Widodo, 2018) membahas tentang klasifikasi kanker payudara dengan metode Random Forest (RF) yang dinilai lebih baik karena memiliki tingkat akurasi yang lebih tinggi dari metode Support Vector Machine (SVM).

Penelitian lainnya dilakukan oleh (Shahura, Soesanto, & Indriani, 2016) memaparkan salah satu cara mendeteksi kanker payudara dengan Fine-Needle Aspiration (FNA) biopsy. Sampel FNA menghasilkan sepuluh karakteristik, yaitu radius, texture, perimeter, area, compactness, smothness, concavity, concabe points, symmetry, dan fractal dimension. Kesepuluh karakteristik tersebut untuk mengklasifikasikan kanker payudara jinak dan ganas dengan menggunakan metode Radial Basis Probabilistic Neural Network (RBPNN).

Sementara pada penelitian ini, dikembangkan suatu sistem clustering dengan menggunakan metode K-Means yang memiliki komponen yaitu 41 data pasien penderita kanker payudara dan 10 gejala pasien yang dijadikan sebagai atribut. Penggunaan metode K-Means clustering diharapkan dapat menghasilkan pengelompokkan yang tepat mengenai tingkat bahaya pada kanker payudara.

Metode K-Means clustering mempartisi data ke dalam cluster/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik yang berbeda dikelompokkan ke dalam cluster yang lain. Adapun tujuan dari clustering yaitu untuk meminimalisasikan variasi di dalam suatu cluster dan memaksimalisasikan variasi antar cluster (Agusta, 2007).

Berdasarkan uraian diatas, dapat diketahui bahwa metode K-Means clustering merupakan salah satu metode yang telah banyak digunakan dalam pembangunan

Volume 10, Nomor 1
p-ISNN 2089-0265
e-ISSN 2598-3016

perangkat lunak untuk diagnosis suatu penyakit.

METODE

Metode yang digunakan dalam penelitian ini adalah K-Means Clustering.

Algoritma K-Means Clustering

Proses perhitungan dari algoritma k-means clustering:

1. Penentuan jumlah cluster
   Penentuan jumlah cluster dilakukan sebagai inisialisasi awal, yang nantinya akan dibagi kedalam data yang ada.
2. Penentuan pusat awal cluster.
   Pada penentuan pusat awal cluster dapat dilakukan dengan mengambil data/nilai secara random, data/nilai tersebut nantinya akan dijadikan pusat awal cluster.
3. Perhitungan pusat jarak cluster. Untuk mengukur jarak dengan pusat cluster menggunakan Rumus Euclidean Distance seperti pada persamaan:

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

4. Pengelompokan Data
   Setelah nilai jarak telah didapatkan. Maka selanjutnya lakukanlah perbandingan dan pilihlah jarak terpendek antara data dengan pusat cluster, kelompokanlah data sesuai letak jarak terpendek dari setiap data.

Data yang digunakan pada penelitian ini diambil dari *Breast Cancer Dataset UCI Machine Learning* sebanyak 569 dengan jumlah pasien kanker payudara *benign* sebanyak 358 data dan jumlah pasien kanker payudara *malignant* sebanyak 212 data, serta terdapat 30 gejala yang terdeteksi oleh pasien kanker payudara.

Data dalam penelitian ini yaitu 41 data pasien yang diambil dari data 70 sampai 110 data, tetapi gejala yang diambil yaitu 10 gejala yang diambil secara acak. Data penelitian dapat dilihat pada tabel 1.

Tabel 1. Data Pasien Kanker Payudara

Id Pasien	Pasien	Radius Mean	Smoothness Mean	Symmetry Mean	Area Se	Fractal Dimension Se	Radius Worst	Smoothness Worst	Symmetry Worst	Area Worst	Fractal Dimension Worst
859487	Pasien 1	12.78	0.09831	0.159	18.33	0.001906	13.46	0.1296	0.2383	554.9	0.0641
859575	Pasien 2	18.94	0.09009	0.1582	96.05	0.001698	24.86	0.1193	0.2551	1866	0.06589
859711	Pasien 3	8.888	0.09783	0.1902	25.44	0.02193	9.733	0.1207	0.2254	284.4	0.1084
859717	Pasien 4	17.2	0.1071	0.1927	69.47	0.006299	23.32	0.1585	0.3313	1681	0.1339
859983	Pasien 5	13.8	0.1007	0.1662	23.35	0.00313	16.57	0.1411	0.2589	812.4	0.103
…	…	…	…	…	…	…	…	…	…	…	…
863270	Pasien 39	12.36	0.08477	0.1602	9.227	0.001356	13.29	0.1184	0.2983	544.1	0.07185
86355	Pasien 40	22.27	0.1326	0.2556	170	0.005037	28.4	0.1701	0.4055	2360	0.09789
864018	Pasien 41	11.34	0.08759	0.1487	16.41	0.002477	13.01	0.1699	0.2829	518.1	0.08832

Volume 10, Nomor 1
p-ISNN 2089-0265
e-ISSN 2598-3016

HASIL DAN PEMBAHASAN

Berdasarkan data pasien kanker payudara yang ditunjukkan pada tabel 1, maka akan dihitung menggunakan metode K-Means Clustering. Pada perhitungan ini akan diberikan parameter-parameter sebagai berikut:

Jumlah cluster :2

Jumlah data :41 (data pasien ke-70 sampai 110)

Jumlah atribut :10 atribut (*Radius Mean, Smoothness Mean, Symmetry Mean, Area Se, Fractal Dimension Se, Radius Worst, Smoothness Worst, Symmetry Worst, Area Worst, dan Fractal Dimention Worst*)

Iterasi 1

1) Penentuan jumlah cluster
   Cluster yang akan digunakan yaitu sejumlah 2 cluster.
2) Penentuan pusat awal cluster.
   Penentuan pusat awal cluster menggunakan nilai MAX dan MIN yang diambil dari tiap-tiap atribut. Nilai MAX dari setiap atribut dijadikan sebagai Pusat Cluster 1 dan nilai MIN dari setiap atribut dijadikan sebagai Pusat Cluster 2. Penentuan pusat awal cluster dapat dilihat pada tabel 2.

Tabel 2. Pusat Awal Cluster

Pusat cluster 1	25.22	0.1398	0.2906	170	0.02193	30	0.1862	0.544	2562	0.1405
Pusat cluster 2	6.981	0.07355	0.135	9.227	0.001356	7.93	0.1006	0.1934	185.2	0.06206

3) Perhitungan pusat jarak cluster.
   Untuk menghitug pusat jarak cluster, gunakan persamaan (1) pada algoritma k-means clustering.

Untuk contoh, perhitungan jarak dari data ke-1 terhadap pusat cluster adalah :

C1=

$$\sqrt{(12.78-25.22)^2+(0.09831-0.1398)^2+(0.159-0.2906)^2+..+(0.0641-0.1405)^2}$$

C2=

$$\sqrt{(12.78-6.981)^2+(0.09831-0.07355)^2+(0.159-0.135)^2+..+(0.0641-0.06206)^2}$$

Dilanjutkan untuk data ke-2 sampai data ke-N. Kemudian akan didapatkan matriks jarak yang ditunjukkan pada tabel 3.

Tabel 3. Hasil Perhitungan Pusat Jarak Cluster

C1	C2	Jarak Terpendek
2012.860945	369.8988618	369.8988618
700.2937767	1683.168596	700.2937767
2282.438676	100.5504672	100.5504672
887.260804	1497.126638	887.260804
1756.028748	627.4555417	627.4555417
…	…	…
…	…	…
2024.45571	358.9803413	358.9803413
202.7073776	2180.884188	202.7073776
2049.799996	333.0447818	333.0447818

Didapatkan data cluster sebagai berikut:

C1: data [2, 4, 9, 10, 14, 17, 19, 27,40]

C2: data [1, 3, 5, 6, 7, 8, 11, 12, 13, 15, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41]

Volume 10, Nomor 1
p-ISNN 2089-0265
e-ISSN 2598-3016

Data di atas jika dikelompokkan maka akan menjadi seperti tabel 4 berikut ini.

Tabel 4. Pengelompokkan Data

No	C1	C2
1		✓
2	✓	
3		✓
4	✓	
5		✓
…	…	…
…	…	…
39		✓
40	✓	
41		✓

Iterasi ke-2

1) Penentuan pusat cluster baru

Setelah mendapatkan hasil perhitungan pusat jarak cluster maka selanjutnya membuat pusat cluster baru. Dimana cara perhitungan berbeda dengan penentuan yang pertama, yaitu dengan menghitung nilai yang sesuai dengan pengelompokkan cluster yang berada di iterasi ke-1. Untuk cluster C1(1,1) nilai didapatkan dari Radius Mean data ke 2, 4, 9, 10, 14, 17, 19, 27, dan 40. Berikut adalah cara perhitungannya:

C1(1,1) = $\frac{x_1+x_2+\cdots+x_n}{n}$

= $\frac{18.94+17.2+18.05+\cdots+22.27}{9}$

= 19.95555556

Untuk cluster C1(1,2) nilai didapatkan dari Smoothness Mean data ke 2, 4, 9, 10, 14, 17, 19, 27, dan 40.

C1(1,2) = $\frac{x_1+x_2+\cdots+x_n}{n}$

= $\frac{0.09831+0.09009+0.09783+\cdots+0.01326}{9}$

= 0.105666667

Lakukan sampai perhitungan C1(1,41). Untuk cluster C2(2,1) nilai didapatkan dari Radius Mean data ke 1, 3, 5, 6, 7, 8, 11, 12, 13, 15, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, dan 41. Berikut adalah cara perhitungannya:

C2(2,1) = $\frac{x_1+x_2+\cdots+x_n}{n}$

= $\frac{12.78+8.888+13.8+12.31+\cdots+11.34}{32}$

= 12.77975

Untuk cluster C2(2,2) nilai didapatkan dari Smoothness Mean data ke 1, 3, 5, 6, 7, 8, 11, 12, 13, 15, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, dan 41.

C2(2,2) = $\frac{x_1+x_2+\cdots+x_n}{n}$

= $\frac{0.09831+0.09783+0.1007+\cdots+0.08759}{32}$

= 0.100474063

Tabel 5. Pusat Cluster Baru Iterasi 2

Cluster Baru	
C1	C2
19.95555556	12.77975
0.105666667	0.100474063
0.212577778	0.1803

Lakukan perhitungan di atas sampai pada C2(2,41), maka akan didapatkan nilai pusat cluster yang baru untuk melakukan perhitungan iterasi ke-2. Setelah pusat cluster baru sudah ditemukan, lakukanlah perhitungan jarak pusat cluster seperti iterasi ke-1. Setelah melakukan perhitungan pada iterasi ke-2 maka akan didapatkan pengelompokkan data cluster sebagai berikut :

C1: data [2, 4, 9, 10, 14, 15, 17, 19, 27,40]
C2: data [1, 3, 5, 6, 7, 8, 11, 12, 13, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, dan 41].

Volume 10, Nomor 1
p-ISNN 2089-0265
e-ISSN 2598-3016

Dilakukan iterasi kembali hingga data tidak mengalami perubahan posisi pada pengelompokkan data dari iterasi sebelumnya. Pada proses perhitungan ini, saat iterasi ke-4 sudah mendapatkan hasil yang konvergen, dimana hasil perhitungan iterasi ke 3 sama dengan yang ke 4. Jadi perhitungan dapat berhenti. Dimana data cluster iterasi ke-empat sebagai berikut :
C1: data [2, 4, 7, 9, 10, 14, 15, 17, 19, 27,40]
C2: data [1, 3, 5, 6, 8, 11, 12, 13, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41],

KESIMPULAN DAN SARAN

Berdasarkan pembahasan di atas maka dapat diambil kesimpulan bahwa 41 data yang telah dilakukan perhitungan menggunakan metode k-means clustering menghasilkan dua cluster dimana cluster ke-1 (C1) terdapat 11 data yaitu data ke-2, 4, 7, 9, 10, 14, 15, 17, 19, 27,40 yang tergolong dalam penyakit kanker payudara ganas. Cluster 2 (C2) terdapat 30 data yaitu data ke-1, 3, 5, 6, 8, 11, 12, 13, 16, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, dan 41. Penulis mengasumsikan bahwa C1 tergolong kanker payudara ganas dan C2 tergolong kanker payudara jinak.

Beberapa saran guna pengembangan diagnosa penyakit kanker payudara agar menjadi lebih baik lagi, yaitu jumlah sampel data harus diperbanyak dan perlu dilakukan uji coba dengan metode pemecahan masalah yang lebih optimal.

DAFTAR PUSTAKA

InfoDATIN. (2016, Oktober). Retrieved Mei 28, 2019, from Pusat Data dan Informasi Kementrian Kesehatan RI: http://www.depkes.go.id/download.php?file=download/pusdatin/infodatin/InfoDatin%20Bulan%20Peduli%20Kanker%20Payudara_2016.pdf

Agusta, Y. (2007). K-Means – Penerapan, Permasalahan dan Metode Terkait. *Jurnal Sistem dan Informatika*, 47-60.

Aliady, H., Tuasikal, N. J., & Widodo, E. (2018). Implementasi Support Vector Machine (SVM) Dan Random Forest Pada Diagnosis Kanker Payudara. SENTIKA 2018, 278-285.

Keleş, A., Keleş, A., & Yavuz, U. (2011). Expert System Based On Neuro-Fuzzy Rules For Diagnosis Breast Cancer. In *Expert Systems With Applications* (pp. 5719-5726). Elsevier.

Shahura, F., Soesanto, O., & Indriani, F. (2016). Penerapan Metode RBPNN Untuk Klasifikasi Kanker Payudara. Kumpulan Jurnal Ilmu Komputer (KLIK), 135-145.