Jurnal Manajemen Dan Administrasi Rumah Sakit (MARSI)
Vol. 10 No 4, Oktober 2026
https://ejournal.urindo.ac.id/index.php/MARSI
E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

KLASTERISASI KARAKTERISTIK PASIEN KANKER PARU DENGAN K-MEANS DI RUMAH SAKIT KANKER DHARMAIS

CLUSTERING OF LUNG CANCER PATIENT CHARACTERISTICS USING K-MEANS AT DHARMAIS CANCER HOSPITAL

Ikmal Aulia Akbar[1], Frestiany Regina Putri[2]
[1,2]Sarjana Terapan Manajemen Informasi Kesehatan, Politeknik Indonusa Surakarta
Email korespondensi: 25.ikmal.akbar@poltekindonusa.ac.id

ABSTRAK

Kanker paru memiliki karakteristik pasien yang heterogen sehingga diperlukan pengelompokan data untuk mendukung perencanaan pelayanan dan peningkatan kualitas registrasi kanker. Penelitian ini bertujuan mengelompokkan karakteristik pasien kanker paru berdasarkan usia, jenis kelamin, topografi, morfologi, dan pekerjaan menggunakan algoritma K-Means. Penelitian menggunakan desain observasional retrospektif terhadap 443 data pasien kanker paru di Rumah Sakit Kanker Dharmais. Nomor rekam medis dan nama pasien tidak digunakan sebagai atribut analisis. Variabel kategorik ditransformasikan menggunakan *one-hot encoding*, sedangkan usia distandardisasi. Pemodelan dilakukan menggunakan Orange Data Mining 3.40 dengan inisialisasi K-Means++, 10 kali pengulangan, dan maksimum 300 iterasi. Jumlah cluster diuji pada K=2 sampai K=10 menggunakan nilai Silhouette. Nilai tertinggi diperoleh pada K=3, yaitu 0,239. Cluster 1 terdiri atas 201 pasien (45,37%) dan seluruhnya laki-laki dengan rerata usia 64,27±6,50 tahun. Cluster 2 terdiri atas 125 pasien (28,22%) dan seluruhnya perempuan dengan rerata usia 59,54±8,40 tahun. Cluster 3 terdiri atas 117 pasien (26,41%), didominasi laki-laki, dan memiliki rerata usia 43,18±7,90 tahun. Adenocarcinoma merupakan morfologi dominan pada seluruh cluster. Topografi C34.9 ditemukan pada lebih dari 90% anggota setiap cluster. Tiga cluster menggambarkan kelompok laki-laki usia lanjut, perempuan usia menengah–lanjut, dan pasien usia lebih muda. Tingginya penggunaan kode topografi tidak spesifik menunjukkan perlunya audit kelengkapan registrasi kanker.

Kata kunci: kanker paru; K-Means; klasterisasi; registrasi kanker; Orange Data Mining

ABSTRACT

Lung cancer patients have heterogeneous characteristics, requiring data segmentation to support service planning and cancer registry quality improvement. This study aimed to cluster lung cancer patients according to age, sex, topography, morphology, and occupation using the K-Means algorithm. A retrospective observational design was employed using 443 lung cancer patient records from Dharmais Cancer Hospital. Medical record numbers and patient names were excluded from the clustering attributes. Categorical variables were transformed using one-hot encoding, while age was standardized. The analysis was performed using Orange Data Mining 3.40 with K-Means++ initialization, 10 reruns, and a maximum of 300 iterations. Candidate cluster numbers from K=2 to K=10 were evaluated using the Silhouette score. The highest score was obtained for K=3 at 0.239. Cluster 1 comprised 201 patients (45.37%), all male, with a mean age of 64.27±6.50 years. Cluster 2 comprised 125 patients (28.22%), all female, with a mean age of 59.54±8.40 years. Cluster 3 comprised 117 patients (26.41%), predominantly male, with a mean age of 43.18±7.90 years. Adenocarcinoma was the dominant morphology across all clusters. Topography code C34.9 accounted for more than 90% of patients in each cluster. The three clusters represented older male patients, middle-to-older-aged female patients, and younger patients. The high proportion of unspecified topography codes indicates the need to audit the completeness of cancer registry data.

Keywords: cancer registry; clustering; K-Means; lung cancer; Orange Data Mining

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

PENDAHULUAN

Kanker paru merupakan masalah kesehatan global dengan angka kejadian dan kematian yang tinggi. World Health Organization memperkirakan terdapat sekitar 2,5 juta kasus baru dan 1,8 juta kematian akibat kanker paru pada tahun 2022 [1]. Estimasi GLOBOCAN juga menunjukkan bahwa kanker paru tetap menjadi salah satu kanker dengan beban tertinggi di dunia [2]. Di Indonesia, estimasi GLOBOCAN 2024 menunjukkan sebanyak 58.446 kasus baru dan 53.323 kematian akibat kanker paru. Kanker paru menempati urutan kedua berdasarkan jumlah kasus baru dan urutan pertama berdasarkan jumlah kematian akibat kanker di Indonesia [3].

Kanker paru tidak merupakan satu kelompok penyakit yang homogen. Perbedaan usia, jenis kelamin, lokasi tumor, morfologi, paparan lingkungan, kebiasaan merokok, dan kondisi klinis dapat membentuk pola pasien yang berbeda [4]. Berdasarkan morfologinya, kanker paru secara umum dibedakan menjadi *non-small cell lung cancer* dan *small cell lung cancer*. Kelompok *non-small cell lung cancer* mencakup adenocarcinoma, squamous cell carcinoma, dan beberapa morfologi lain [4,5].

Data registrasi kanker yang lengkap dan terstandardisasi diperlukan untuk menggambarkan beban penyakit, mendukung perencanaan pelayanan, mengevaluasi hasil pelayanan, dan mengembangkan kebijakan pengendalian kanker. Kualitas informasi tersebut dipengaruhi oleh kelengkapan, validitas, konsistensi, komparabilitas, dan ketepatan waktu data [6]. Dalam praktiknya, data registrasi rumah sakit sering kali hanya disajikan dalam bentuk distribusi frekuensi sehingga hubungan antarkarakteristik pasien belum tergambarkan secara optimal.

Klasterisasi dapat digunakan untuk mengidentifikasi kelompok alami dalam data tanpa memerlukan label kelas sebelumnya. K-Means merupakan salah satu algoritma klasterisasi yang banyak digunakan karena relatif sederhana, efisien, dan mudah diinterpretasikan [7]. Meskipun K-Means pada dasarnya dirancang untuk data numerik, variabel kategorik dapat ditransformasikan menjadi variabel indikator sebelum perhitungan jarak [8]. Orange Data Mining menyediakan lingkungan pemrograman visual yang mendukung praproses, pemodelan, evaluasi, dan visualisasi hasil klasterisasi [9].

Penelitian ini bertujuan mengelompokkan karakteristik pasien kanker paru di Rumah Sakit Kanker Dharmais berdasarkan usia, jenis kelamin, topografi, morfologi, dan pekerjaan menggunakan algoritma K-Means. Hasil penelitian diharapkan dapat memberikan profil kelompok pasien untuk mendukung pengelolaan registrasi kanker, perencanaan pelayanan, dan evaluasi kualitas data rumah sakit.

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

TINJAUAN PUSTAKA

Karakteristik Kanker Paru

Kanker paru terdiri atas beberapa kelompok morfologi dengan karakteristik biologis dan klinis yang berbeda. *Non-small cell lung cancer* mencakup sekitar 85% kasus kanker paru, sedangkan *small cell lung cancer* lebih jarang tetapi cenderung memiliki perkembangan yang lebih agresif [1,4]. Adenocarcinoma merupakan salah satu morfologi yang paling banyak ditemukan, termasuk pada perempuan dan pasien yang tidak memiliki riwayat merokok. Sementara itu, squamous cell carcinoma dan small cell carcinoma memiliki hubungan yang lebih kuat dengan paparan tembakau [5,7].

Topografi kanker paru diklasifikasikan berdasarkan lokasi primer tumor menggunakan *International Classification of Diseases for Oncology*. Kode C34 meliputi bronkus utama, lobus atas, lobus tengah, lobus bawah, lesi tumpang tindih, serta bronkus atau paru yang tidak ditentukan secara spesifik. Penggunaan kode yang terlalu umum dapat mengurangi ketepatan analisis epidemiologi maupun perencanaan pelayanan.

Algoritma K-Means

K-Means membagi objek ke dalam sejumlah K cluster dengan meminimalkan jumlah kuadrat jarak antara setiap objek dan pusat cluster. Proses tersebut meliputi pemilihan pusat awal, pengalokasian objek ke pusat terdekat, perhitungan ulang pusat cluster, dan pengulangan sampai tidak terjadi perubahan yang berarti [7].

Pemilihan pusat awal menggunakan K-Means++ dapat mengurangi kemungkinan konvergensi pada solusi lokal yang kurang baik [10]. Jumlah cluster dapat dievaluasi menggunakan nilai Silhouette. Nilai Silhouette membandingkan kedekatan suatu objek dengan anggota cluster yang sama dan jaraknya terhadap cluster terdekat lainnya [11]. Nilai yang semakin mendekati satu menunjukkan pemisahan cluster yang semakin baik.

METODE PENELITIAN

Penelitian ini merupakan penelitian kuantitatif dengan desain observasional retrospektif. Data penelitian berasal dari data sekunder pasien kanker paru di Rumah Sakit Kanker Dharmais.

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

Populasi penelitian adalah seluruh data pasien kanker paru yang tersedia dalam periode penelitian. Pengambilan sampel menggunakan *total sampling*. Setelah proses pemeriksaan kelengkapan dan pembersihan data, sebanyak 443 data pasien digunakan dalam analisis.

Kriteria inklusi meliputi data pasien kanker paru yang memiliki informasi usia, jenis kelamin, topografi, morfologi, dan pekerjaan. Data dengan atribut penelitian yang tidak lengkap atau mengalami duplikasi dikeluarkan dari analisis. Nomor rekam medis dan nama pasien hanya digunakan untuk pemeriksaan data dan tidak digunakan sebagai atribut pembentuk cluster.

Variabel yang digunakan terdiri atas:

1. usia dalam tahun;
2. jenis kelamin;
3. topografi kanker paru;
4. morfologi kanker paru; dan
5. pekerjaan.

Usia juga dikelompokkan menjadi 18–45 tahun, 45–59 tahun, dan ≥60 tahun untuk keperluan interpretasi profil cluster. Kode topografi terdiri atas C34.0, C34.1, C34.2, C34.3, C34.8, dan C34.9. Morfologi dikelompokkan menjadi adenocarcinoma, carcinoma squamous cell, large cell carcinoma, non-small cell carcinoma, dan small cell carcinoma.

Analisis dilakukan menggunakan Orange Data Mining versi 3.40. Variabel kategorik diubah melalui *one-hot encoding*, sedangkan usia distandardisasi menjadi rerata nol dan varians satu. Standardisasi dilakukan agar rentang usia tidak mendominasi perhitungan jarak. Proses analisis menggunakan rangkaian widget:

File → Continuize → K-Means → Data Table/Distributions/Box Plot.

K-Means dijalankan menggunakan inisialisasi K-Means++, 10 kali pengulangan, dan maksimum 300 iterasi. Kandidat jumlah cluster diuji dari K=2 sampai K=10. Jumlah cluster akhir dipilih berdasarkan nilai rata-rata Silhouette tertinggi. Hasil klasterisasi disajikan dalam bentuk jumlah, persentase, rerata, standar deviasi, median, dan rentang antarkuartil. Perbedaan usia antarcluster diperiksa menggunakan analisis varians satu arah. Nilai $p<0,05$ dianggap bermakna.

HASIL DAN PEMBAHASAN

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

Karakteristik Pasien

Sebanyak 443 data pasien memenuhi kriteria analisis. Rerata usia pasien berdasarkan pembobotan rerata setiap cluster adalah sekitar 57,37 tahun. Sebanyak 208 pasien (46,95%) berusia ≥60 tahun dan 173 pasien (39,05%) berusia 45–59 tahun. Pasien laki-laki berjumlah 295 orang (66,59%), sedangkan pasien perempuan berjumlah 148 orang (33,41%).

Tabel 1. Karakteristik pasien kanker paru

Karakteristik	n	%
Kategori usia		
18–45 tahun	62	14,00
45–59 tahun	173	39,05
≥60 tahun	208	46,95
Jenis kelamin		
Laki-laki	295	66,59
Perempuan	148	33,41
Topografi		
C34.0—Bronkus utama	2	0,45
C34.1—Lobus atas	22	4,97
C34.2—Lobus tengah	1	0,23
C34.3—Lobus bawah	7	1,58
C34.8—Lesi tumpang tindih	1	0,23
C34.9—Bronkus atau paru tidak spesifik	410	92,55
Morfologi		
Adenocarcinoma	292	65,91
Carcinoma squamous cell	74	16,70
Large cell carcinoma	2	0,45
Non-small cell carcinoma	29	6,55
Small cell carcinoma	46	10,38

Adenocarcinoma merupakan morfologi terbanyak, yaitu 292 pasien atau 65,91%. Carcinoma squamous cell ditemukan pada 74 pasien (16,70%), small cell carcinoma pada 46 pasien (10,38%), non-small cell carcinoma yang tidak dirinci pada 29 pasien (6,55%), dan large cell carcinoma pada dua pasien (0,45%). Sebanyak 410 pasien atau 92,55% menggunakan kode topografi C34.9. Dominasi kode tersebut menunjukkan bahwa lokasi anatomis tumor yang lebih rinci belum banyak tercatat dalam dataset.

Penentuan Jumlah Cluster

Pengujian K=2 sampai K=10 menunjukkan nilai Silhouette tertinggi pada K=3, yaitu 0,239. Oleh karena itu, pemodelan akhir menggunakan tiga cluster. Nilai Silhouette yang relatif rendah menunjukkan bahwa masih terdapat kedekatan atau tumpang tindih karakteristik antarcluster. Hasil klasterisasi perlu dipahami sebagai segmentasi eksploratif dan tidak digunakan sebagai sistem diagnosis maupun klasifikasi klinis.

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

Cluster 1 merupakan kelompok terbesar dengan 201 pasien (45,37%). Cluster 2 terdiri atas 125 pasien (28,22%), sedangkan Cluster 3 terdiri atas 117 pasien (26,41%).

Tabel 2. Distribusi usia dan jenis kelamin berdasarkan cluster

Karakteristik	C1, n=201	C2, n=125	C3, n=117
Proporsi dari seluruh pasien	45,37%	28,22%	26,41%
Rerata usia±SD, tahun	64,27±6,50	59,54±8,40	43,18±7,90
Median usia	64	59	44
Kuartil 1–Kuartil 3	59–68	53–65	39–49
Kategori usia			
18–45 tahun	0 (0,00%)	3 (2,40%)	59 (50,43%)
45–59 tahun	52 (25,87%)	63 (50,40%)	58 (49,57%)
≥60 tahun	149 (74,13%)	59 (47,20%)	0 (0,00%)
Jenis kelamin			
Laki-laki	201 (100%)	0 (0,00%)	94 (80,34%)
Perempuan	0 (0,00%)	125 (100%)	23 (19,66%)

Keterangan: persentase dihitung terhadap jumlah anggota masing-masing cluster.

Hasil analisis varians menunjukkan adanya perbedaan usia yang bermakna antarcluster dengan nilai F=302,630 dan p<0,001. Namun, karena usia merupakan salah satu atribut yang digunakan dalam pembentukan cluster, hasil tersebut terutama menggambarkan keberhasilan algoritma dalam memisahkan kelompok berdasarkan usia dan tidak dapat ditafsirkan sebagai hubungan sebab-akibat.

Karakteristik Topografi dan Morfologi

Tabel 3. Distribusi topografi dan morfologi berdasarkan cluster

Karakteristik	C1, n=201	C2, n=125	C3, n=117
Topografi			
C34.0	2 (1,00%)	0 (0,00%)	0 (0,00%)
C34.1	10 (4,98%)	8 (6,40%)	4 (3,42%)
C34.2	1 (0,50%)	0 (0,00%)	0 (0,00%)
C34.3	1 (0,50%)	4 (3,20%)	2 (1,71%)
C34.8	1 (0,50%)	0 (0,00%)	0 (0,00%)
C34.9	186 (92,54%)	113 (90,40%)	111 (94,87%)
Morfologi			
Adenocarcinoma	104 (51,74%)	103 (82,40%)	85 (72,65%)
Carcinoma squamous cell	43 (21,39%)	13 (10,40%)	18 (15,38%)
Large cell carcinoma	2 (1,00%)	0 (0,00%)	0 (0,00%)
Non-small cell carcinoma	15 (7,46%)	5 (4,00%)	9 (7,69%)
Small cell carcinoma	37 (18,41%)	4 (3,20%)	5 (4,27%)

Keterangan: persentase dihitung terhadap jumlah anggota masing-masing cluster.

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

Kode topografi C34.9 mendominasi seluruh cluster, yaitu 92,54% pada C1, 90,40% pada C2, dan 94,87% pada C3. Distribusi tersebut relatif homogen sehingga topografi tidak menjadi atribut pembeda utama antarcluster.

Adenocarcinoma menjadi morfologi dominan pada semua cluster. Proporsinya paling besar pada C2, yaitu 82,40%, diikuti C3 sebesar 72,65% dan C1 sebesar 51,74%. Sebaliknya, C1 mempunyai proporsi carcinoma squamous cell dan small cell carcinoma yang lebih tinggi dibandingkan cluster lainnya.

Karakteristik Pekerjaan

Tabel 4. Lima pekerjaan dengan frekuensi tertinggi pada setiap cluster

Cluster	Pekerjaan	n	%
C1	Karyawan swasta	52	25,87
	Wiraswasta	50	24,88
	Pensiunan	22	10,95
	Buruh harian lepas	21	10,45
	Belum/tidak bekerja	15	7,46
C2	Mengurus rumah tangga	86	68,80
	Karyawan swasta	10	8,00
	Pensiunan	7	5,60
	Pegawai Negeri Sipil	7	5,60
	Belum/tidak bekerja	3	2,40
C3	Karyawan swasta	58	49,57
	Wiraswasta	23	19,66
	Mengurus rumah tangga	9	7,69
	Belum/tidak bekerja	7	5,98
	Buruh harian lepas	6	5,13

Keterangan: kategori pekerjaan lain tetap dimasukkan dalam pembentukan cluster, tetapi hanya lima kategori terbanyak yang ditampilkan.

Profil Cluster 1: Laki-Laki Usia Lanjut

Cluster 1 terdiri atas 201 pasien dan seluruh anggotanya berjenis kelamin laki-laki. Rerata usia pasien adalah 64,27 tahun dan 74,13% berusia ≥60 tahun. Cluster ini juga memiliki proporsi small cell carcinoma sebesar 18,41% dan carcinoma squamous cell sebesar 21,39%, lebih tinggi dibandingkan dua cluster lainnya.

Secara epidemiologis, kanker paru lebih banyak ditemukan pada usia lanjut karena proses karsinogenesis berhubungan dengan akumulasi paparan dalam jangka panjang [7,12]. Small cell carcinoma dan squamous cell carcinoma juga dilaporkan memiliki hubungan kuat dengan paparan rokok [4,7].

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

Meskipun demikian, penelitian ini tidak memiliki data kebiasaan merokok sehingga tingginya kedua morfologi pada C1 tidak dapat secara langsung dihubungkan dengan paparan tembakau.

Pekerjaan terbanyak pada C1 adalah karyawan swasta dan wiraswasta. Informasi pekerjaan pada penelitian ini menggambarkan status pekerjaan pada saat data dicatat, tetapi belum mencakup riwayat pekerjaan, durasi kerja, maupun jenis paparan pekerjaan. Oleh karena itu, pekerjaan tidak dapat ditafsirkan sebagai faktor risiko tanpa data paparan yang lebih rinci.

Profil Cluster 2: Perempuan Usia Menengah hingga Lanjut

Cluster 2 terdiri atas 125 pasien dan seluruh anggotanya berjenis kelamin perempuan. Rerata usia pasien adalah 59,54 tahun. Sebanyak 50,40% pasien berusia 45–59 tahun dan 47,20% berusia ≥60 tahun.

Adenocarcinoma ditemukan pada 82,40% anggota C2. Temuan ini sejalan dengan literatur yang menunjukkan bahwa adenocarcinoma merupakan morfologi kanker paru yang banyak ditemukan pada perempuan dan juga menjadi morfologi utama pada pasien yang tidak merokok [4,5,13]. Namun, tidak tersedianya informasi status merokok, paparan asap rokok, polusi udara, penggunaan bahan bakar rumah tangga, serta faktor genetik menyebabkan pola tersebut tidak dapat digunakan untuk menyimpulkan penyebab penyakit.

Sebagian besar anggota C2 mempunyai pekerjaan mengurus rumah tangga. Dominasi kategori tersebut berkaitan dengan komposisi jenis kelamin dalam cluster. Untuk analisis faktor risiko, data pekerjaan sebaiknya dilengkapi dengan informasi aktivitas rumah tangga, jenis bahan bakar memasak, ventilasi, paparan asap, serta durasi paparan.

Profil Cluster 3: Pasien Usia Lebih Muda

Cluster 3 memiliki rerata usia paling rendah, yaitu 43,18 tahun. Sebanyak 50,43% anggota C3 berusia 18–45 tahun dan tidak terdapat pasien berusia ≥60 tahun. Cluster ini didominasi laki-laki sebesar 80,34%, tetapi masih mencakup 19,66% perempuan.

Adenocarcinoma ditemukan pada 72,65% anggota C3. Karyawan swasta dan wiraswasta merupakan pekerjaan yang paling banyak ditemukan. Karakteristik tersebut menunjukkan adanya kelompok pasien kanker paru usia produktif. Kelompok ini dapat memiliki kebutuhan pelayanan yang berbeda, seperti penyesuaian jadwal pelayanan, dukungan untuk mempertahankan fungsi kerja, rehabilitasi, konseling psikososial, dan perencanaan pembiayaan jangka panjang.

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

Temuan pasien kanker paru pada usia lebih muda juga memperlihatkan perlunya pengumpulan faktor risiko yang lebih komprehensif. Selain merokok, data registrasi dapat mencakup paparan asap rokok, polusi udara, radon, bahan kimia di tempat kerja, riwayat keluarga, serta riwayat penyakit paru kronis [1,7].

Implikasi terhadap Manajemen dan Administrasi Rumah Sakit

Hasil klasterisasi memberikan segmentasi awal yang dapat digunakan untuk membantu perencanaan pelayanan. C1 menggambarkan kelompok laki-laki usia lanjut yang kemungkinan membutuhkan koordinasi pelayanan geriatri, penilaian komorbiditas, dan dukungan pelayanan kanker yang lebih kompleks. C2 menggambarkan kelompok perempuan usia menengah–lanjut dengan dominasi adenocarcinoma. C3 menggambarkan kelompok usia produktif yang dapat memerlukan dukungan pelayanan terkait pekerjaan, keluarga, dan keberlanjutan pengobatan.

Hasil paling penting dari aspek manajemen informasi kesehatan adalah tingginya penggunaan kode topografi C34.9. Penggunaan kode tidak spesifik pada lebih dari 90% pasien di semua cluster membatasi kemampuan rumah sakit untuk mengevaluasi pola lokasi primer tumor. Kualitas data registrasi kanker seharusnya mencakup validitas, komparabilitas, kelengkapan, dan ketepatan waktu [6,14,15].

Perbaikan dapat dilakukan melalui penetapan elemen lokasi tumor sebagai data wajib, integrasi laporan radiologi dan patologi, pemeriksaan konsistensi kode topografi dan morfologi, serta pemberian umpan balik kepada petugas registrasi. Audit berkala terhadap penggunaan kode tidak spesifik juga diperlukan untuk membedakan kasus yang memang tidak dapat ditentukan lokasinya dan kasus yang belum dikodekan secara lengkap.

Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan. Pertama, data berasal dari satu rumah sakit sehingga profil cluster belum dapat digeneralisasikan ke seluruh pasien kanker paru di Indonesia. Kedua, nilai Silhouette sebesar 0,239 menunjukkan pemisahan cluster yang masih terbatas. Ketiga, variabel yang dianalisis belum mencakup kebiasaan merokok, stadium, komorbiditas, hasil pemeriksaan molekuler, terapi, status pembiayaan, dan luaran pasien.

Keempat, K-Means menggunakan jarak Euclidean dan pada dasarnya lebih sesuai untuk data numerik. Transformasi *one-hot encoding* memungkinkan variabel kategorik dianalisis, tetapi hasilnya tetap dipengaruhi oleh cara pengkodean atribut [8,16]. Penelitian berikutnya dapat membandingkan hasil K-

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

Means dengan K-Prototypes, *partitioning around medoids* menggunakan jarak Gower, atau metode klasterisasi lain yang dirancang untuk data campuran. Kelima, label cluster bersifat eksploratif dan bukan kategori diagnosis atau tingkat risiko klinis.

KESIMPULAN DAN SARAN

Algoritma K-Means menghasilkan tiga cluster karakteristik pasien kanker paru dengan nilai Silhouette tertinggi sebesar 0,239. Cluster 1 terdiri atas 201 pasien dan menggambarkan laki-laki usia lanjut dengan proporsi squamous cell carcinoma serta small cell carcinoma yang lebih tinggi. Cluster 2 terdiri atas 125 pasien dan menggambarkan perempuan usia menengah–lanjut dengan dominasi adenocarcinoma. Cluster 3 terdiri atas 117 pasien dan menggambarkan kelompok usia lebih muda yang didominasi laki-laki.

Usia dan jenis kelamin menjadi atribut pembeda paling jelas, sedangkan topografi relatif homogen karena lebih dari 90% pasien pada setiap cluster menggunakan kode C34.9. Adenocarcinoma menjadi morfologi dominan pada seluruh cluster.

Rumah sakit disarankan melakukan audit terhadap penggunaan kode topografi tidak spesifik, meningkatkan integrasi data klinis dengan registrasi kanker, dan melengkapi variabel status merokok, paparan lingkungan, stadium, terapi, serta luaran pasien. Hasil cluster dapat digunakan sebagai informasi pendukung perencanaan pelayanan, tetapi tidak digunakan sebagai dasar diagnosis atau keputusan klinis individual. Validasi pada data rumah sakit lain dan perbandingan dengan metode klasterisasi untuk data campuran diperlukan pada penelitian selanjutnya.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Rumah Sakit Kanker Dharmais yang telah mendukung pelaksanaan penelitian dan pengelolaan data penelitian.

KONFLIK KEPENTINGAN

Penulis menyatakan tidak terdapat konflik kepentingan dalam penelitian dan publikasi artikel ini.

DAFTAR PUSTAKA

[1]. World Health Organization. Lung cancer. World Health Organization, Geneva, 2026.

E-ISSN: 2865-6583
P-ISSN: 2868-6298
10.52643/marsi.v10i4.9178

[2]. Sung H, Filho AM, Laversanne M, Ferlay J, Siegel RL, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. CA Cancer J Clin. 2026. doi:10.3322/caac.70090.

[3]. Ervik M, Laversanne M, Lam F, Colombet M, Ferlay J, Filho AM, Bray F. Global Cancer Observatory: Cancer Today, 2024 Estimates, Version 1.0. International Agency for Research on Cancer, Lyon, 2026.

[4]. Thai AA, Solomon BJ, Sequist LV, Gainor JF, Heist RS. Lung cancer. Lancet. 2021;398(10299):535–554. doi:10.1016/S0140-6736(21)00312-3.

[5]. WHO Classification of Tumours Editorial Board. Thoracic Tumours. 5th ed. International Agency for Research on Cancer, Lyon, 2021.

[6]. [6] Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer. 2009;45(5):747–755. doi:10.1016/j.ejca.2008.11.032.

[7]. Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inf Sci. 2023;622:178–210. doi:10.1016/j.ins.2022.11.139.

[8]. Ahmad A, Dey L. A K-mean clustering algorithm for mixed numeric and categorical data. Data Knowl Eng. 2007;63(2):503–527. doi:10.1016/j.datak.2007.03.016.

[9]. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, et al. Orange: data mining toolbox in Python. J Mach Learn Res. 2013;14(71):2349–2353.

[10]. Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, 2007, p.1027–1035.

[11]. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987; 20:53–65. doi:10.1016/0377-0427(87)90125-7.

[12]. Barta JA, Powell CA, Wisnivesky JP. Global epidemiology of lung cancer. Ann Glob Health. 2019;85(1):8. doi:10.5334/aogh.2419.

[13]. Bade BC, Dela Cruz CS. Lung cancer 2020: epidemiology, etiology, and prevention. Clin Chest Med. 2020;41(1):1–24. doi:10.1016/j.ccm.2019.10.001.

[14]. Parkin DM, Bray F. Evaluation of data quality in the cancer registry: principles and methods. Part II: completeness. Eur J Cancer. 2009;45(5):756–764. doi:10.1016/j.ejca.2008.11.033.

[15]. Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, Whelan S, editors. International Classification of Diseases for Oncology. 3rd ed. World Health Organization, Geneva, 2013.

[16]. Huang Z. Extensions to the K-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov. 1998; 2:283–304. doi:10.1023/A:1009769707641.