Implementasi Algoritma *K-Means* Dalam Pengelompokan Kasus Penyakit Tuberkulosis Paru Berdasarkan Provinsi

Vivi Febriyanti[1], Heru Satria Tambunan[2], Ilham Syahputra Saragih[3], Irfan Sudahri Damanik[4], Harly Okprana[5]
STIKOM Tunas Bangsa, Pematangsiantar, Indonesia
Jl. Jenderal Sudirman Blok A No 1-3 Pematangsiantar, Indonesia
E-mail : vivifebriyanti18@gmail.com

Abstract

Pulmonary tuberculosis is a lung disease caused by germs that cause symptoms of excessive coughing. In Indonesia, pulmonary tuberculosis is a disease in the top five countries with many pulmonary tuberculosis. The purpose of this study was to determine the high and low number of cases of pulmonary tuberculosis in the province. In this study the data used were sourced from the National Statistics Agency for 2007-2015. For this reason the authors use data mining techniques in the data processing with k-means clustering method to obtain information based on data that is processed as a reference to find out the number of cases of pulmonary tuberculosis that most suffered by province. The results of this study are grouping the number of cases of pulmonary tuberculosis with 3 clusters, namely high cluster, medium cluster, low cluster. From the calculation of k-means, there were 3 provinces as high clusters, 3 provinces as medium clusters, and 28 provinces as low clusters. The implementation process using the RapidMiner 5.3 application is used to help find accurate values.

Keywords: *Pulmonary tuberculosis, Data Mining, K-Means, RapidMiner,BPS*

1. Pendahuluan

Penyakit tuberkulosis paru merupakan penyakit menular yang menyerang anak-anak maupun orang dewasa yang disebabkan oleh bakteri yang keluar dari orang yang mengidap penyakit tuberkulosis paru. Bakteri yang dikeluarkan biasanya dominan menyerang paru-paru manusia sehingga disebut dengan tuberkulosis paru. Gejala penyakit tuberkulosis paru ditandai dengan batuk yang berlebihan (dalam waktu 3 minggu), batuk yang berdahak bahkan mengeluarkan darah. Di Indonesia penyakit tuberkulosis paru merupakan penyakit yang ditakuti karena menyebabkan kematian pada anak-anak maupun orang dewasa. Secara umum penyakit tuberkulosis paru disebabkan oleh beberapa faktor yaitu lingkungan, cuaca, tempat tinggal dan keturunan. Penyebaran penyakit tuberkulosis paru sudah menyebar luas khususnya wilayah Indonesia. Data Mining merupakan cara untuk menemukan informasi yang tersembunyi dalam sebuah basis data dan merupakan bagian dari proses Knowledge Discovery in Database (KDD) untuk menemukan informasi dan pola yang berguna dalam data [1]. Untuk bisa mewujudkan hal ini penulis memerlukan data yang akurat untuk melakukan pengelompokan data jumlah kasus penyakit tuberkulosis paru. Rekapitulasi data pada BPS Provinsi setiap tahunnya mengalami penuruan dan peningkatan terkait penyakit tuberkulosis paru, sehingga perlu dilakukannya penelitian menggunakan *k-means clustering* untuk mengetahui *cluster* tinggi, *cluster* sedang, *cluster* rendah. Data diperoleh dari BPS Provinsi mengenai data jumlah kasus penyakit tuberkulosis paru berdasarkan provinsi di Indonesia pada tahun 2007-2015.

K-means merupakan suatu algoritma yang digunakan dalam pengelompokkan secara pertisi yang memisahkan data ke dalam kelompok yang berbeda-berda. Algoritma ini mampu meminimalkan jarak antara data ke clusternya [2]. Metode ini 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

ISSN: 2686-0260

yang berbeda dikelompokkan ke dalam kelompok yang lain [3]. Untuk mengetahui jumlah kasus penyakit tuberkulosis paru suatu provinsi di Indonesia penulis menggunakan teknik *data mining* dalam proses pengolahan data dengan algoritma *k-means.* Data yang telah didapatkan selanjutnya dikelompokkan kedalam *cluster* dan diuji menggunakan *RapidMiner* versi 5.3. Berdasarkan permasalahan di atas, maka penulis mengangkat judul skripsi "**Implementasi Algoritma *K-Means* Dalam Pengelompokan Kasus Penyakit Tuberkulosis Paru Berdasarkan Provinsi".** Diharapkan dengan adanya penelitian ini dapat memberikan informasi kepada pemerintah mengenai *cluster* kasus penyakit tuberkulosis paru disetiap provinsi melalui kegiatan sosialisasi dan penanganan cepat tanggap terhadap kasus penyakit tuberkulosis paru agar mengurangi jumlah kasus penyakit tuberkulosis paru untuk tahun-tahun berikutnya.

2. Metodologi Penelitian

Tujuan dari skripsi ini adalah mengumpulkan informasi dan mengelola data untuk menyelesaikan permasalahan dalam penelitian untuk mencari jawaban terhadap suatu masalah yang akan diteliti. Dengan menggunakan teknik data mining yaitu algoritma *k-means clustering*.

2.1. Data Mining

Data mining merupakan sebuah inti dari proses KDD, meliputi dugaan algoritma yang mengeksplor data, membangun model dan menemukan pola yang belum diketahui [4].

2.2. *K-Means*

K-Means merupakan salah satu metode pengelompokan data nonhierarki (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok [5].

2.3. *Clustering*

Clustering ialah teknik data mining yang digunakan untuk menganalisis dan mengkaji data untuk menyelesaikan permasalahan dalam pengelompokkan data membagi dari suatu dataset ke dalam subset [6].

2.4. *RapidMiner*

Rapid Miner merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari *Institute of Technologi Blanchardstown* dan Ralf Klinkenberg dari rapid-i.com dengan tampilan GUI *(Graphical User Interface)* sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini [7].

3. Analisa Dan Pembahasan

3.1. Metode Pengumpulan Data

Penelitian ini dilakukan di Indonesia dengan pengambilan data langsung dari Badan Pusat Statistik (BPS) Nasional dengan situs resmi bps.go.id (https://www.bps.go.id/site/resultTab) dan waktu pengumpulan data dilakukan selama 2 minggu yaitu dari tanggal 14 Oktober s/d 27 Oktober 2019.

3.2. Tahap Pengolahan Data

Data yang telah diperolah akan diolah terlebih dahulu untuk dapat di *clustering.* Dalam tahap sebelumnya, data setiap provinsi akan dijumlah setiap aspeknya sehingga pada tahapan ini sudah diperoleh perhitungan nilai yang akan diproses pada tahap *clustering* [8].

3.3. Tahap *Clustering*

Dalam metode *clustering* konsep utama yang ditekankan adalah pencarian pusat *cluste*r secara iteratif, dimana pusat *cluster* ditentukan berdasarkan jarak minimum setiap data pada pusat *cluster* [9].

3.4. Tahap Analisis

Pada tahap analisis dilakukan analisa data terhadap jumlah kasus penyakit tuberkulosis paru berdasarkan provinsi. Pada tahap sebelumya sudah ditentukan menjadi 3 *cluster* yaitu *cluster* tinggi, *cluster* sedang, dan *cluster* rendah. Pada tahap inilah akan dilakukan anilisis hasil. Dalam metode *k-means clustering* terlebih dahulu dilakukan pencarian rata-rata terhadap jumlah kasus penyakit tuberkulosis paru berdasarkan provinsi. Hasil rata-rata yang sudah diakumulasikan dapat dilihat pada tabel dibawah ini :

Tabel 1. Nilai Rata-Rata Penyakit Tuberkulosis Paru

No	Provinsi	Rata-Rata	No	Provinsi	Rata-Rata
1	Aceh	4091	18	Nusa Tenggara Barat	5260
2	Sumatera Utara	18108	19	Nusa Tenggara Timur	5097
3	Sumatera Barat	5772	20	Kalimantan Barat	5162
4	Riau	4258	21	Kalimantan Tengah	2022
5	Jambi	3179	22	Kalimantan Selatan	4482
6	Sumatera Selatan	7562	23	Kalimantan Timur	3454
7	Bengkulu	1762	24	Kalimantan Utara	87
8	Lampung	7106	25	Sulawesi Utara	5189
9	Kep. Bangka Belitung	1241	26	Sulawesi Tengah	2827
10	Kep. Riau	1720	27	Sulawesi Selatan	8692
11	Dki Jakarta	21363	28	Sulawesi Tenggara	3440
12	Jawa Barat	54366	29	Gorontalo	1630
13	Jawa Tengah	32905	30	Sulawesi Barat	1324
14	Di Yogyakarta	2176	31	Maluku	2879
15	Jawa Timur	35920	32	Maluku Utara	1126
16	Banten	13270	33	Papua Barat	1835
17	Bali	2746	34	Papua	5063

Setelah data sudah diakumulasikan dan di cari nilai rata-ratanya, selanjutnya menentukan nilai k jumlah *cluster* adapun *cluster* yang dibentuk yaitu *cluster* tinggi, *cluster* sedang, *cluster* rendah.

3.5. *Centroid* Data

Menentukan nilai centroid (pusat *cluster*) awal yang telah ditentukan secara random berdasarkan nilai variabel data yang di *cluster* sebanyak yang telah ditentukan. *Cluster* tinggi (c1) diperoleh dari nilai maximum pada rata-rata data, *cluster* sedang (c2) diperoleh dari nilai rata-rata, dan *cluster* rendah (c3) diperoleh dari nilai minimum pada rata-rata data. Berikut adalah nilai *centroid* data awal untuk iterasi 1:

Tabel 2. *Centroid* Data Awal (Iterasi 1)

C1 = Maximum	54366
C2 = Average	8150
C3 = Minimum	87

3.6. *Clustering* Data

Menghitung jarak setiap data penyakit tuberkulosis paru tehadap pusat *cluster*. Setelah data nilai pusat *cluster* ditentukan, maka langkah selanjutnya adalah menghitung

jarak masing-masing data terhadap pusat *cluster* dengan menggunakan rumus sebagai berikut:

Dilakukan perhitungan jarak terhadap data penyakit tuberkulosis paru dengan titik pusat *(centroid)* pada *cluster* pertama:

D (1.1) = $\sqrt{(54366-4091)^2}$ = **50275**

D (1.2) = $\sqrt{(54366-18108)^2}$ = **36258**

D (1.3) = $\sqrt{(54366-5772)^2}$ = **48594**

Lakukan perhitungan yang sama sampai D (1.34).

Berikut tabel 3 hasil perhitungan jarak data dengan titik pusat iterasi 1 menggunakan *Elulidean Distance*.

Tabel 3. Perhitungan Jarak Pusat *Custer* Iterasi 1

No	Provinsi	C1	C2	C3	Jarak Terpendek
1	Aceh	50275	4059	4004	4004
2	Sumatera Utara	36258	9958	18021	9958
3	Sumatera Barat	48594	2378	5685	2378
4	Riau	50108	3892	4171	3892
5	Jambi	51187	4971	3092	3092
6	Sumatera Selatan	46804	588	7475	588
7	Bengkulu	52604	6388	1675	1675
8	Lampung	47260	1044	7019	1044
9	Kep. Bangka Belitung	53125	6909	1154	1154
10	Kep. Riau	52646	6430	1633	1633
11	Dki Jakarta	33003	13213	21276	13213
12	Jawa Barat	0	46216	54279	0
13	Jawa Tengah	21461	24755	32818	21461
14	Di Yogyakarta	52190	5974	2089	2089
15	Jawa Timur	18446	27770	35833	18446
16	Banten	41096	5120	13183	5120
17	Bali	51620	5404	2659	2659
18	Nusa Tenggara Barat	49106	2890	5173	2890
19	Nusa Tenggara Timur	49269	3053	5010	3053
20	Kalimantan Barat	49204	2988	5075	2988
21	Kalimantan Tengah	52344	6128	1935	1935
22	Kalimantan Selatan	49884	3668	4395	3668
23	Kalimantan Timur	50912	4696	3367	3367
24	Kalimantan Utara	54279	8063	0	0
25	Sulawesi Utara	49177	2961	5102	2961
26	Sulawesi Tengah	51539	5323	2740	2740
27	Sulawesi Selatan	45674	542	8605	542
28	Sulawesi Tenggara	50926	4710	3353	3353
29	Gorontalo	52736	6520	1543	1543
30	Sulawesi Barat	53042	6826	1237	1237
31	Maluku	51487	5271	2792	2792
32	Maluku Utara	53240	7024	1039	1039
33	Papua Barat	52531	6315	1748	1748
34	Papua	49303	3087	4976	3087

Tabel 4. Hasil Pengelompokan Iterasi 1

No	Provinsi	C1	C2	C3	No		C1	C2	C3
1	Aceh			1	18	Nusa Tenggara Barat		1	
2	Sumatera Utara		1		19	Nusa Tenggara Timur		1	
3	Sumatera Barat		1		20	Kalimantan Barat		1	
4	Riau		1		21	Kalimantan Tengah			1
5	Jambi			1	22	Kalimantan Selatan		1	
6	Sumatera Selatan		1		23	Kalimantan Timur			1
7	Bengkulu			1	24	Kalimantan Utara			1
8	Lampung		1		25	Sulawesi Utara		1	
9	Kep. Bangka Belitung			1	26	Sulawesi Tengah			1
10	Kep. Riau			1	27	Sulawesi Selatan		1	
11	Dki Jakarta		1		28	Sulawesi Tenggara			1
12	Jawa Barat	1			29	Gorontalo			1
13	Jawa Tengah	1			30	Sulawesi Barat			1
14	Di Yogyakarta			1	31	Maluku			1
15	Jawa Timur	1			32	Maluku Utara			1
16	Banten		1		33	Papua Barat			1
17	Bali			1	34	Papua		1	

Tabel 5. Hasil *Cluster* Iterasi 1

Cluster	Hasil
C1	3
C2	14
C3	17

Selanjutnya dilakukan kembali langkah 4-5. Jika nilai *centroid* hasil iterasi pertama dengan nilai sebelumnya bernilai sama ataupun nilai *centroid* sudah optimal serta posisi *cluster* pada data tuberkulosis paru tidak mengalami perubahan lagi maka poses iterasi berhenti. Namun jika posisi *cluster* belum sama maka proses iterasi masih berlanjut pada iterasi berikutnya sampai *cluster* bernilai sama.Prioses iterasi berhenti pada iterasi ke 6. Berikut hasil pengelompokan data iterasi ke 6 dapaat dilihat pada tabel dibawah ini :

Tabel 6. *Centroid* Data Iterasi 6

No	Provinsi	C1	C2	C3	Jarak Terpendek
1	Aceh	36973	13489	477	477
2	Sumatera Utara	22956	528	14494	528
3	Sumatera Barat	35292	11808	2158	2158
4	Riau	36806	13322	644	644
5	Jambi	37885	14401	435	435
6	Sumatera Selatan	33502	10018	3948	3948
7	Bengkulu	39302	15818	1852	1852
8	Lampung	33958	10474	3492	3492
9	Kep. Bangka Belitung	39823	16339	2373	2373
10	Kep. Riau	39344	15860	1894	1894
11	Dki Jakarta	19701	3783	17749	3783
12	Jawa Barat	13302	36786	50752	13302
13	Jawa Tengah	8159	15325	29291	8159
14	Di Yogyakarta	38888	15404	1438	1438
15	Jawa Timur	5144	18340	32306	5144
16	Banten	27794	4310	9656	4310
17	Bali	38318	14834	868	868
18	Nusa Tenggara Barat	35804	12320	1646	1646
19	Nusa Tenggara Timur	35967	12483	1483	1483
20	Kalimantan Barat	35902	12418	1548	1548

No	Provinsi	C1	C2	C3	Jarak Terpendek
21	Kalimantan Tengah	39042	15558	1592	1592
22	Kalimantan Selatan	36582	13098	868	868
23	Kalimantan Timur	37610	14126	160	160
24	Kalimantan Utara	40977	17493	3527	3527
25	Sulawesi Utara	35875	12391	1575	1575
26	Sulawesi Tengah	38237	14753	787	787
27	Sulawesi Selatan	32372	8888	5078	5078
28	Sulawesi Tenggara	37624	14140	174	174
29	Gorontalo	39434	15950	1984	1984
30	Sulawesi Barat	39740	16256	2290	2290
31	Maluku	38185	14701	735	735
32	Maluku Utara	39938	16454	2488	2488
33	Papua Barat	39229	15745	1779	1779
34	Papua	36001	12517	1449	1449

Tabel 7. Hasil Pengelompokan Iterasi 1

No	Provinsi	C1	C2	C3	No		C1	C2	C3
1	Aceh			1	18	Nusa Tenggara Barat			1
2	Sumatera Utara		1		19	Nusa Tenggara Timur			1
3	Sumatera Barat			1	20	Kalimantan Barat			1
4	Riau			1	21	Kalimantan Tengah			1
5	Jambi			1	22	Kalimantan Selatan			1
6	Sumatera Selatan			1	23	Kalimantan Timur			1
7	Bengkulu			1	24	Kalimantan Utara			1
8	Lampung			1	25	Sulawesi Utara			1
9	Kep. Bangka Belitung			1	26	Sulawesi Tengah			1
10	Kep. Riau			1	27	Sulawesi Selatan			1
11	Dki Jakarta		1		28	Sulawesi Tenggara			1
12	Jawa Barat	1			29	Gorontalo			1
13	Jawa Tengah	1			30	Sulawesi Barat			1
14	Di Yogyakarta			1	31	Maluku			1
15	Jawa Timur	1			32	Maluku Utara			1
16	Banten		1		33	Papua Barat			1
17	Bali			1	34	Papua			1

Tabel 8. Hasil *Cluster* Iterasi 6

Cluster	Hasil
C1	3
C2	3
C3	28

3.7. Analisa Data

Perhitungan manual pada data hasil penyakit tuberkulosis paru diatas didapatkan sebuah hasil akhir yang sama dimana pada iterasi ke 5 dan ke 6 didapatkan hasil yang sama. Hasil dari kedua iterasi tersebut bernilai C1 = 3, C2 = 3, dan C3 = 28 pada posisi tiap *cluster* x sehingga posisi *cluster* pada data tersebut tidak mengalami perubahan lagi maka proses iterasi berhenti. Berdasarkaan posisi *cluster* masing-masing data penyakit tuberkulosis paru dan nilai *cluster* hasil ietrasi keenam.

4. Kesimpulan

Hasil akhir penelitian dari jumlah keseluruhan provinsi disimpulkan bahwa telah di dapat nilai dengan 3 *cluster* yaitu *cluster* tinggi, *cluster* sedang, *cluster* rendah dalam pengelompokan terhadap jumlah kasus penyakit tuberkulosis paru yaitu :

a) (C1) *Cluster* tinggi dengan jumlah data penyakit tuberkulosis paru sebanyak 3 provinsi yaitu, Jawa Barat, Jawa Tengah, Jawa Timur.
b) (C2) *Cluster* sedang dengan jumlah data penyakit tuberkulosis paru sebanyak 3 provinsi yaitu, Sumatera Utara, DKI Jakarta, Banten.
c) (C3) *Cluster* rendah dengan jumlah data penyakit tuberkulosis paru sebanyak 28 provinsi yaitu, Aceh, Sumater Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep Bangka Belitung, Kep Riau, DI Yogyakarta, Bali, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua Barat, Papua.

Penelitian ini dapat dijadikan bahan perbandingan untuk peneliti lain yang ingin mengangkat judul yang sama guna mengetahui hasil dari nilai penelitian untuk tahun-tahun berikutnya sehingga bisa dijadikan masukan kepada pihak pemerintah khususnya terhadap jumlah kasus penyakit tuberkulosis paru sebagai bahan pertimbangan untuk *cluster* tinggi dalam upaya menurunkan jumlah kasus penyakit tuberkulosis paru pada tahun-tahun berikutnya. Pada penelitian selanjutnya lebih baik menggunakan data yang terbaru mengenai jumlah kasus penyakit tuberkulosis paru berdasarkan provinsi.

Daftar Pustaka

[1] F. Kurnia, I. Fahmi, E. Wahyudi, and G. E. S. Mige, "PENERAPAN ALGORITMA K-MEANS UNTUK PENGELOMPOKAN DIAGNOSA PENYAKIT MATA BERDASARKAN RENTANG USIA," vol. 2, no. 1, 2019.
[2] R. W. Sari, D. Hartama, I. Gunawan, and P. Windarto, "Aplikasi RapidMiner dalam Pengelompokkan Kasus Penyakit AIDS berdasarkan Provinsi dengan Data Mining K-means Clustering," pp. 59–69.
[3] P. D. P. Silitonga and I. S. Morina, "Klusterisasi Pola Penyebaran Penyakit Pasien Berdasarkan Usia Pasien Dengan Menggunakan K-Means Clustering," vol. VI, no. 2, pp. 2005–2008, 2017.
[4] F. Hardiyanti, H. S. Tambunan, and I. S. Saragih, "PENERAPAN METODE K-MEDOIDS CLUSTERING PADA PENANGANAN KASUS DIARE DI INDONESIA," vol. 3, no. 2012, pp. 598–603, 2019.
[5] M. G. Sadewo, A. P. Windarto, and D. Hartama, "PENERAPAN DATAMINING PADA POPULASI DAGING AYAM RAS PEDAGING DI INDONESIA BERDASARKAN PROVINSI MENGGUNAKAN K-MEANS," pp. 60–67, 2016.
[6] P. Alkhairi and A. P. Windarto, "Penerapan K-Means Cluster Pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara," pp. 762–767, 2019.
[7] S. Haryati, A. Sudarsono, and E. Suryana, "IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA C4.5 (STUDI KASUS: UNIVERSITAS DEHASEN BENGKULU)," vol. 11, no. 2, pp. 130–138, 2015.
[8] R. W. Sari, A. Wanto, and A. P. Windarto, "IMPLEMENTASI RAPIDMINER DENGAN METODE K-MEANS (STUDY KASUS: IMUNISASI CAMPAK PADA BALITA BERDASARKAN PROVINSI)," vol. 2, pp. 224–230, 2018.
[9] Karmila, H. S. Tambunan, Sumarno, and A. P. Windarto, "Penerapan Data Mining K-Means dalam Mengelompokkan Kasus Penyakit Malaria Berdasarkan Provinsi dengan Aplikasi RapidMiner," pp. 31–40, 2018.