Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

Penerapan Algoritma *K-Means Clustering* Dalam Persentase Merokok Pada Penduduk Umur Di Atas 15 Tahun Menurut Provinsi

Rahmat Kurniawan[1], Suhada[2], Rafiqa Dewi[2]

[1]Sistem Informasi, STIKOM Tunas Bangsa, Pematangsiantar, Indonesia
[2]AMIK Tunas Bangsa, Pematangsiantar, Indonesia
Email: [1]rahmat16s04@gmail.com, [2]suhada@amiktunasbangsa.ac.id, [3]rafiqa@amiktunasbangsa.ac.id
Submitted: **25/12/2020**; Accepted: **15/01/2021**; Published: **24/01/2021**

Abstrak– Penelitian ini bertujuan untuk mengetahui banyaknya penduduk yang merokok usia diatas 15 Tahun dengan menerapkan Data Mining. Penerapannya menggunakan Algoritma *K-means* dimana data yang dimasukkan adalah persentase merokok. Dengan menggunakan metode ini data-data yang di peroleh dapat dikelompokkan menjadi 2 klaster. Penelitian ini menggunakan data sekunder yaitu data diperoleh dari media perantara yang terekam di situs Badan Pusat Statistik dengan alamat url https://www.bps.go.id/. Hasil yang didapat Dalam penelitian ini adalah pengelompokkan jumlah persentase merokok yang di kelompokkan sebanyak 2 klaster yaitu klaster 1 (klaster tertinggi) mendapatkan hasil 22 provinsi dan klaster 2 (klaster terendah) mendapatkan 12 provinsi.

Kata Kunci: Data Mining; *Clustering*, *K-means*; Persentase Merokok; RapidMiner

Abstract–This research aims to determine the number of people who are smoking the age of 15 years by applying Data Mining. The application uses the K-means algorithm where the data entered is the percentage of smoking. Using this method the data obtained can be grouped into 2 klasters. This research uses secondary data that is obtained from intermediate media that is recorded on the site of the Central Statistics website with the URL address https://www.bps.go.id/. The results gained in this study are grouping the number of smoking percentages that are grouped by as many as 2 klasters is klaster 1 (highest klaster) to get results 22 provinces and klaster 2 (lowest klaster) get 12 provinces.

Keywords: Data Mining; Clustering; K-means; Percentage of Smoking; RapidMiner

1. PENDAHULUAN

Di era globalisasi, teknologi merupakan salah satu solusi bagi manusia dalam menyelesaikan suatu masalah. Hal ini merupakan faktor pendorong adanya algoritma-algoritma yang muncul dari Data Mining. Salah satu algoritma dari data mining ini adalah Algoritma *K-means* yang merupakan metode dalam teknik *Clustering*. *K-means* merupakan salah satu metode pengelompokkan data nonhierarki (sekatan) yang berusaha mempartisi data yang ada kedalam bentuk dua atau lebih kelompok.

Indonesia merupakan negara ketiga dengan jumlah perokok terbesar setalah China dan India. Merokok merupakan kebiasaan buruk yang sudah dianggap biasa, karena begitu banyak nya para perokok atau juga karena begitu banyaknya aktifitas merokok yang biasa dijumpai disekitar kita sehingga merokok menjadi hal yang lumrah dan biasa saja.

Data dari Badan Pusat Statistik (BPS) menyebutkan bahwa di Indonesia pada tahun 2015 persentase merokok pada penduduk umur di atas 15 tahun sebesar 30,08% dan mengalami penurunan pada tahun 2016 menjadi 28,97%. Pada tahun 2017 persentase merokok mengalami peningkatan dari tahun sebelumnya menjadi 29,25% dan pada tahun 2018 terus meningkat sebesar 32,20%. Dari data tersebut penulis ingin mengelompokkan provinsi mana yang persentase merokok umur di atas 15 tahun yang tinggi dan rendah. Diharapkan penelitian ini dapat mengetahui klaster persentasi merokok pada tingkat tinggi dan rendah sehingga dapat menjadi masukkan kepada pemerintah melalui kegiatan "Hari Tanpa Tembakau Sedunia" agar dapat mengurangi angka persentase merokok disetiap wilayah.

2. METODE PENELITIAN

2.1 Data Mining

Data mining merupakan serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Data mining mulai ada sejak 1990-an sebagai cara yang benar dan tepat untuk mengambil pola dan informasi yang digunakan untuk menemukan hubungan antara data untuk melakukan pengelompokkan ke dalam satu atau lebih cluster sehingga objek - objek yang berada dalam satu cluster akan mempunyai kesamaan yang tinggi antara satu dengan lainnya. Data mining merupakan bagian dari proses penemuan pengetahuan dari basis data Knowledge Discovery in [1], [2].

2.2 RapidMiner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (open source). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

keputusan yang paling baik. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan menggunakan bahasa java sehingga dapat bekerja di semua sistem operasi.[3]

2.3 Metode *K-Means*

K-Means merupakan salah satu metode data clustering non hierarki yang berusaha mempartisi data yang ada dalam bentuk satu atau lebih cluster, sehingga data dengan karakteristik yang sama dikelompokkan dalam satu cluster yang sama pula. Data dengan karakteristik yang berbeda dikelompokkan dalam kelompok yang lain. Metode yang termasuk dalam algoritma clustering berbasis jarak yang membagi data kedalam sejumlah cluster dan algoritma ini hanya bekerja pada atribut numerik [4]–[6].
Langkah-langkah algoritma *K-means* [7]

1. Menentukan k sebagai jumlah *klaster* yang ingin dibentuk.
2. Mengalokasikan data ke dalam *klaster* secara acak.
3. Menentukan pusat *klaster* (*centroid*) dari data yang ada pada masing-masing *klaster*.
   dengan pesamaan :
   $C_{kj} = \frac{x_{1j}+x_{2j}+\cdots+x_{nj}}{n}$ (1)
   dimana
   C_{kj}= pusat *klaster* ke-k pada variabel ke j (j = 1,2,...,p)
   n = banyak data pada *klaster* ke-k
4. Menentukan jarak setiap objek dengan setiap *centroid* dengan perhitungan jarak setiap objek dengan setiap *centroid* menggunakan jarak *Euclidean.*
   $d(X_i, X_g) = \sqrt{\sum_{j=1}^{p}(X_{ij} - X_{gj})^2}$ (2)
5. Menghitung fungsi objektif dengan formula:
   $J = \sum_{i=1}^{n}\sum_{j=1}^{k} a_{ij} d(x_i, C_{kj})^2$ (3)
6. Mengalokasikan masing-masing data ke *centroid*/rata-rata terdekat yang dirumuskan sebagai berikut:
   $a_{ij} = \begin{cases} 1, & s = min\{d(x_i, C_{kj})\} \\ 0, & lainnya \end{cases}$ (4)
   a_{ij} adalah nilai keanggotaan titik x_i ke pusat *klaster* C_{kj}, s adalah jarak terpendek dari data x_i ke pusat *klaster* C_{kj} setelah dibandingkan.
7. Mengulangi kembali langkah 3-6 sampai tidak ada lagi perpindahan objek atau tidak ada perubahan pada fungsi objektifnya.

3. HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian adalah data persentase merokok umur diatas 15 tahun menurut provinsi dari tahun 2015-2018. Kumpulan data yang diperoleh penulis digunakan sebagai data masukan dalam membuat model aturan menggunakan algoritma *K-means* dan menggunakan *software Rapidminer.*

3.1 Analisa Data

Analisis data yang digunakan penelitian ini menggunakan data kuantitatif dengan teknik analisis data yang menggunakan jenis *statistic deskriptif*. *Statistik deskriptif* adalah menganalisis data dengan cara mendeskripsikan atau menggambarkan data yang telah terkumpul tanpa bermaksud membuat kesimpulan yang berlaku untuk umum atau generalisasi. Data yang diperoleh kemudian diolah dengan *rapidminer* menggunakan *performance* yang berfungsi sebagai validasi dan reabilitas data untuk mencari keakuratan data.

Tabel 1. Data Persentase Merokok Pada Penduduk Umur ≥ 15 Tahun
(Sumber :Badan Pusat Statistik)

Provinsi	Persentase Merokok Pada Penduduk Umur ≥ 15 Tahun			
	2015	2016	2017	2018
Aceh	29.82	28.16	28.85	32.20
Sumatera Utara	29.15	27.88	28.47	31.76
Sumatera Barat	32.41	30.59	31.71	31.10
Riau	31.21	29.61	29.34	35.32
Jambi	30.82	29.18	29.18	32.72
Sumatera Selatan	33.13	31.57	32.46	28.21
Bengkulu	33.68	33.15	33.41	33.07

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

Provinsi	Persentase Merokok Pada Penduduk Umur ≥ 15 Tahun			
	2015	2016	2017	2018
Lampung	34.12	33.39	33.75	35.53
Kep. Bangka Belitung	30.70	29.32	29.67	35.95
Kep. Riau	29.18	29.25	29.98	32.32
DKI Jakarta	27.31	26.42	24.72	29.67
Jawa Barat	33.82	32.67	33.19	30.77
Jawa Tengah	28.57	27.19	27.69	35.78
DI Yogyakarta	24.12	23.11	22.92	30.79
Jawa Timur	29.03	28.16	27.69	25.80
Banten	32.95	31.64	31.77	30.66
Bali	22.96	21.62	22.22	34.93
Nusa Tenggara Barat	31.60	30.88	30.59	26.05
Nusa Tenggara Timur	25.47	24.91	27.31	33.92
Kalimantan Barat	29.35	28.09	28.84	31.30
Kalimantan Tengah	30.53	29.21	29.24	30.92
Kalimantan Selatan	25.76	25.34	25.03	32.64
Kalimantan Timur	25.59	25.23	24.69	27.18
Kalimantan Utara	28.61	28.38	28.18	29.17
Sulawesi Utara	29.31	29.23	29.27	29.82
Sulawesi Tengah	32.56	31.88	32.18	32.80
Sulawesi Selatan	25.49	25.13	25.44	35.57
Sulawesi Tenggara	28.49	27.60	29.22	29.51
Gorontalo	33.93	31.71	34.46	31.46
Sulawesi Barat	28.29	27.36	26.59	36.56
Maluku	27.19	25.68	27.46	29.41
Maluku Utara	31.14	30.23	30.57	32.74
Papua Barat	29.28	26.18	27.60	35.29
Papua	26.67	24.04	27.28	32.73

3.2 Penerapan Algoritma K-Means

Berikut ini merupakan langkah-langkah yang dilakukan penulis dalam mengelompokkan menggunakan Algoritma *K-means Clustering*:

1) Menentukan data yang ingin di klaster. Data yang digunakan pada proses *Clustering* yaitu Data persentase merokok diatas 15 tahun menurut provinsi dari tahun 2015-2018 dengan menggunakan data sebanyak 34 provinsi. Pada proses *Clustering* diawali dengan mencari nilai rata-rata untuk setiap provinsi. Berikut ini beberapa contoh mencari nilai rata – rata :

   R1 = 29,82 + 28,16 + 28,85 + 32,2 / 4 = 29,7575
   R2 = 29,15 + 27,88 + 28,47 + 31,76 / 4 = 29,315
   R3 = 32,41 + 30,59 + 31,71 + 31,1 / 4 = 31,4525
   R4 = 31,21 + 29,61 + 29,34 + 35,32 / 4 = 31,37

   Perhitungan mencari nilai rata-rata dilanjutkan sampai R34 dan Hasilnya dapat dilihat pada tabel dibawah ini:

   Iterasi 1

Tabel 2. Jumlah Rata-Rata Setiap Provinsi

No	Provinsi	Rata-rata
1	Aceh	29,7575
2	Sumatera Utara	29,315
3	Sumatera Barat	31,4525
4	Riau	31,37
5	Jambi	30,475
6	Sumatera Selatan	31,3425
7	Bengkulu	33,3275
8	Lampung	34,1975
9	Kep. Bangka Belitung	31,41
10	Kep. Riau	30,1825
11	DKI Jakarta	27,03
12	Jawa Barat	32,6125
13	Jawa Tengah	29,8075
14	DI Yogyakarta	24,12

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

No	Provinsi	Rata-rata
15	Jawa Timur	27,67
16	Banten	31,755
17	Bali	25,4325
18	Nusa Tenggara Barat	29,78
19	Nusa Tenggara Timur	27,9025
20	Kalimantan Barat	29,395
21	Kalimantan Tengah	29,975
22	Kalimantan Selatan	27,1925
23	Kalimantan Timur	25,6725
24	Kalimantan Utara	28,585
25	Sulawesi Utara	29,4075
26	Sulawesi Tengah	32,355
27	Sulawesi Selatan	27,9075
28	Sulawesi Tenggara	28,705
29	Gorontalo	32,89
30	Sulawesi Barat	29,7
31	Maluku	27,435
32	Maluku Utara	31,17
33	Papua Barat	29,5875
34	Papua	27,68

2) Mengembangkan nilai k sebagai pusat klaster awal (centroid) persentase merokok umur diatas 15 tahun sebanyak 2 klaster. Adapun klaster yang dibentuk yaitu klaster tinggi (C1) dan klaster rendah (C2). Klaster tinggi (C1) diperoleh dari nilai tertinggi yang terdapat pada tabel 2 dan klaster rendah diperoleh dari nilai terendah pada tabel 2. Cara mencari nilai centroid awal untuk literasi 1 yaitu:

C1 = Max (29,7575; 29,315; 31,4525; 31,37; 30,475; 31,3425; 33,3275; 34,1975; 31,41; 30,1825; 27,03; 32,6125; 29,8075; 24,12; 27,67; 31,755; 25,4325; 29,78; 27,9025; 29,395; 29,975; 27,1925; 25,6725; 28,585; 29,4075; 32,355; 27,9075; 28,705; 32,89; 29,7; 27,435; 31,17; 29,5875; 27,68) = 34,1975

C2 = Min (29,7575; 29,315; 31,4525; 31,37; 30,475; 31,3425; 33,3275; 34,1975; 31,41; 30,1825; 27,03; 32,6125; 29,8075; 24,12; 27,67; 31,755; 25,4325; 29,78; 27,9025; 29,395; 29,975; 27,1925; 25,6725; 28,585; 29,4075; 32,355; 27,9075; 28,705; 32,89; 29,7; 27,435; 31,17; 29,5875; 27,68) = 24,12

Berikut ini merupakan hasil dari centroid data awal :

Tabel 3. Centroid Data Awal

C1	34,1975
C2	24,12

3) Menghitung jarak setiap jumlah data persentase merokok terhadap masing-masing klaster sehingga ditemukan jarak terdekat dari setiap data dengan cetroid. Berikut ini contoh perhitungan jarak setiap data pada klaster pertama (C1) :

D(1.1) = $\sqrt{(29{,}7575 - 34{,}1975)^2}$ = 4,44

D(1.2) = $\sqrt{(29{,}315 - 34{,}1975)^2}$ = 4,8825

D(1.3) = $\sqrt{(31{,}4525 - 34{,}1975)^2}$ = 2,745

D(1.4) = $\sqrt{(31{,}37 - 34{,}1975)^2}$ = 2,8275

Sampai dengan D(1.34)

Perhitungan jarak pada setiap data pada klaster kedua (C2) :

D(2.1) = $\sqrt{(29{,}7575 - 24{,}12)^2}$ = 5,6375

D(2.2) = $\sqrt{(29{,}315 - 24{,}12)^2}$ = 5,195

D(2.3) = $\sqrt{(31{,}4525 - 24{,}12)^2}$ = 7,3325

D(2.4) = $\sqrt{(31{,}37 - 24{,}12)^2}$ = 7,25

Sampai dengan D(2.34)

Berikut tabel 4. hasil perhitungan jarak data dengan titik pusat pada Literasi 1:

Tabel 4. Hasil Perhitungan Jarak Data Dengan Titik Pusat Klaster

No	Provinsi	Rata-rata	C1	C2	Jarak Terpendek
1	Aceh	29,7575	4,44	5,6375	4,44
2	Sumatera Utara	29,315	4,8825	5,195	4,8825

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

3	Sumatera Barat	31,4525	2,745	7,3325	2,745
4	Riau	31,37	2,8275	7,25	2,8275
5	Jambi	30,475	3,7225	6,355	3,7225
6	Sumatera Selatan	31,3425	2,855	7,2225	2,855
7	Bengkulu	33,3275	0,87	9,2075	0,87
8	Lampung	34,1975	0	10,0775	0
9	Kep. Bangka Belitung	31,41	2,7875	7,29	2,7875
10	Kep. Riau	30,1825	4,015	6,0625	4,015
11	DKI Jakarta	27,03	7,1675	2,91	2,91
12	Jawa Barat	32,6125	1,585	8,4925	1,585
13	Jawa Tengah	29,8075	4,39	5,6875	4,39
14	DI Yogyakarta	24,12	10,0775	0	0
15	Jawa Timur	27,67	6,5275	3,55	3,55
16	Banten	31,755	2,4425	7,635	2,4425
17	Bali	25,4325	8,765	1,3125	1,3125
18	Nusa Tenggara Barat	29,78	4,4175	5,66	4,4175
19	Nusa Tenggara Timur	27,9025	6,295	3,7825	3,7825
20	Kalimantan Barat	29,395	4,8025	5,275	4,8025
21	Kalimantan Tengah	29,975	4,2225	5,855	4,2225
22	Kalimantan Selatan	27,1925	7,005	3,0725	3,0725
23	Kalimantan Timur	25,6725	8,525	1,5525	1,5525
24	Kalimantan Utara	28,585	5,6125	4,465	4,465
25	Sulawesi Utara	29,4075	4,79	5,2875	4,79
26	Sulawesi Tengah	32,355	1,8425	8,235	1,8425
27	Sulawesi Selatan	29,9075	6,29	3,7875	3,7875
28	Sulawesi Tenggara	28,705	5,4925	4,585	4,585
29	Gorontalo	32,89	1,3075	8,77	1,3075
30	Sulawesi Barat	29,7	4,4975	5,58	4,4975
31	Maluku	27,435	6,7625	3,315	3,315
32	Maluku Utara	31,17	3,0275	7,05	3,0275
33	Papua Barat	29,5875	4,61	5,4675	4,61
34	Papua	27,68	6,5175	3,56	3,56

4) Mengelompokkan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil). Jika nilai terendah terdapat di klaster 1 (C1) maka masuk kedalam kelompok klaster 1 dan begitu juga sebaliknya, dapat dilihat pada tabel 5. dan hasilnya sebagai berikut:

Tabel 5. Hasil Klaster Literasi 1

Klaster	Hasil
C1	23
C2	11

5) Setelah hasil klaster pada literasi 1 di dapat langkah selanjutnya yaitu memperbaruhi nilai centroid baru.
6) Mengulang langkah 2 hingga 5 sampai anggota tiap klaster tidak ada yang berubah.
Contoh perhitungan mencari nilai centroid baru untuk Literasi 2 yaitu :

$$C1 = \frac{\begin{array}{c}29{,}7575 + 29{,}315 + 31{,}4525 + 31{,}37 + 30{,}475 \\ + 31{,}3425 + 33{,}3275 + 34{,}1975 + 31{,}41 + 30{,}1825 \\ + 32{,}6125 + 29{,}8075 + 31{,}755 + 29{,}78 + 29{,}395 \\ + 29{,}975 + 27{,}1925 + 29{,}4075 + 32{,}355 + 32{,}89 + 29{,}7 + 31{,}17 + 29{,}5875\end{array}}{23} = 30{,}803$$

$$C2 = \frac{\begin{array}{c}27{,}03 + 24{,}12 + 27{,}67 + 25{,}4325 + 27{,}9025 + \\ 25{,}6725 + 28{,}585 + 27{,}9075 + 28{,}705 + 27{,}435 + 27{,}68\end{array}}{11} = 27{,}104$$

Tabel 6. Hasil Perhitungan Centroid baru pada Literasi 2

Klaster	Nilai
C1	30,803
C2	27,104

Setelah mendapatkan titik pusat klaster (centroid) baru maka langkah selanjutnya yaitu menghitung kembali jarak setiap data persentase merokok terhadap masing-masing pusat klaster. Perhitungan jarak setiap data pada klaster pertama(C1) :

$$E(1.1) = \sqrt{(29{,}7575 - 30{,}803)^2} = 1{,}045$$

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

E(1.2) = $\sqrt{(29{,}315 - 30{,}803)^2}$ = 1,4875
E(1.3) = $\sqrt{(31{,}4525 - 30{,}803)^2}$ = 0,65
E (1.4) = $\sqrt{(31{,}37 - 30{,}803)^2}$ = 0,5675
Sampai dengan E(1.34)

Perhitungan jarak pada setiap data pada klaster kedua(C2) :

E(2.1) = $\sqrt{(29{,}7575 - 27{,}104)^2}$ = 2,653864
E (2.2) = $\sqrt{(29{,}315 - 27{,}104)^2}$ = 2,211364
E (2.3) = $\sqrt{(31{,}4525 - 27{,}104)^2}$ = 4,348864
E (2.4) = $\sqrt{(31{,}37 - 27{,}104)^2}$ = 4,266364
Sampai dengan E(2.34)

Berikut tabel 7 hasil perhitungan jarak data dengan titik pusat pada literasi 2.

Tabel 7. Perhitungan Jarak Data Literasi 2

No	Provinsi	Rata-rata	C1	C2	Jarak Terpendek
1	Aceh	29,7575	1,045	2,653864	1,045
2	Sumatera Utara	29,315	1,4875	2,211364	1,4875
3	Sumatera Barat	31,4525	0,65	4,348864	0,65
4	Riau	31,37	0,5675	4,266364	0,5675
5	Jambi	30,475	0,3275	3,371364	0,3275
6	Sumatera Selatan	31,3425	0,54	4,238864	0,54
7	Bengkulu	33,3275	2,525	6,223864	2,525
8	Lampung	34,1975	3,395	7,093864	3,395
9	Kep. Bangka Belitung	31,41	0,6075	4,306364	0,6075
10	Kep. Riau	30,1825	0,62	3,078864	0,62
11	DKI Jakarta	27,03	3,7725	0,073636	0,073636364
12	Jawa Barat	32,6125	1,81	5,508864	1,81
13	Jawa Tengah	29,8075	0,995	2,703864	0,995
14	DI Yogyakarta	24,12	6,6825	2,983636	2,983636364
15	Jawa Timur	27,67	3,1325	0,566364	0,566363636
16	Banten	31,755	0,9525	4,651364	0,9525
17	Bali	25,4325	5,37	1,671136	1,671136364
18	Nusa Tenggara Barat	29,78	1,0225	2,676364	1,0225
19	Nusa Tenggara Timur	27,9025	2,9	0,798864	0,798863636
20	Kalimantan Barat	29,395	1,4075	2,291364	1,4075
21	Kalimantan Tengah	29,975	0,8275	2,871364	0,8275
22	Kalimantan Selatan	27,1925	3,61	0,088864	0,088863636
23	Kalimantan Timur	25,6725	5,13	1,431136	1,431136364
24	Kalimantan Utara	28,585	2,2175	1,481364	1,481363636
25	Sulawesi Utara	29,4075	1,395	2,303864	1,395
26	Sulawesi Tengah	32,355	1,5525	5,251364	1,5525
27	Sulawesi Selatan	27,9075	2,895	0,803864	0,803863636
28	Sulawesi Tenggara	28,705	2,0975	1,601364	1,601363636
29	Gorontalo	32,89	2,0875	5,786364	2,0875
30	Sulawesi Barat	29,7	1,1025	2,596364	1,1025
31	Maluku	27,435	3,3675	0,331364	0,331363636
32	Maluku Utara	31,17	0,3675	4,066364	0,3675
33	Papua Barat	29,5875	1,215	2,483864	1,215
34	Papua	27,68	3,1225	0,576364	0,576363636

Hasil perhitungan pada data klaster literasi 2 yaitu:

Tabel 8. Hasil Klaster Literasi 2

Klaster	Hasil
C1	22
C2	12

Setelah hasil pada literasi 2 di dapatkan, bandingkan hasil pada literasi 1 dan literasi 2, jika tidak ada perubahan maka proses berhenti, jika ada perubahan maka proses akan dilanjutkan kembali hingga mendapatkan hasil yang sama. Dan hasil yang sama dengan literasi 1 adalah literasi 3.
Berikut contoh perhitungan mencari nilai centroid baru untuk Literasi 3 yaitu :

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

$$C1 = \frac{\begin{matrix}29{,}7575 + 29{,}315 + 31{,}4525 + 31{,}37 + 30{,}475 \\ + 31{,}3425 + 33{,}3275 + 34{,}1975 + 31{,}41 + 30{,}1825 \\ + 32{,}6125 + 29{,}8075 + 31{,}755 + 29{,}78 + 29{,}395 \\ + 29{,}975 + 29{,}4075 + 32{,}355 + 32{,}89 + 29{,}7 + 31{,}17 + 29{,}5875\end{matrix}}{22} = 30{,}967$$

$$C2 = \frac{\begin{matrix}27{,}03 + 24{,}12 + 27{,}67 + 25{,}4325 + 27{,}9025 + 27{,}1925 + \\ 25{,}6725 + 28{,}585 + 27{,}9075 + 28{,}705 + 27{,}435 + 27{,}68\end{matrix}}{12} = 25{,}026$$

Tabel 9. Hasil Perhitungan Centroid baru pada Literasi 3

Klaster	Nilai
C1	30,967
C2	25,026

Setelah mendapatkan titik pusat klaster (centroid) baru maka langkah selanjutnya yaitu menghitung kembali jarak setiap data jumlah persentase merokok terhadap masing-masing pusat klaster. Perhitungan jarak setiap data pada klaster pertama (C1) :

F(1.1) $=\sqrt{(29{,}7575 - 30{,}967)^2}$ = 1,209090909
F(1.2) $=\sqrt{(29{,}315 - 30{,}967)^2}$ = 1,651590909
F(1.3) $=\sqrt{(31{,}4525 - 30{,}967)^2}$ = 0,485909091
F (1.4) $=\sqrt{(31{,}37 - 30{,}967)^2}$ = 0,403409091
Sampai dengan F(1.34)

Perhitungan jarak pada setiap data pada klaster kedua (K2) :

F(2.1) $=\sqrt{(29{,}7575 - 25{,}026)^2}$ = 4,7319
F(2.2) $=\sqrt{(29{,}315 - 25{,}026)^2}$ = 4,2894
F(2.3) $=\sqrt{(31{,}4525 - 25{,}026)^2}$ = 6,4269
F(2.4) $=\sqrt{(31{,}37 - 25{,}026)^2}$ = 6,3444
Sampai dengan F(2.34)

Berikut hasil perhitungan jarak data dengan titik pusat pada literasi 3.

Tabel 10. Perhitungan Jarak Data Literasi 3

No	Provinsi	Rata-rata	C1	C2	Jarak Terpendek
1	Aceh	29,7575	1,209090909	2,646458	1,209090909
2	Sumatera Utara	29,315	1,651590909	2,203958	1,651590909
3	Sumatera Barat	31,4525	0,485909091	4,341458	0,485909091
4	Riau	31,37	0,403409091	4,258958	0,403409091
5	Jambi	30,475	0,491590909	3,363958	0,491590909
6	Sumatera Selatan	31,3425	0,375909091	4,231458	0,375909091
7	Bengkulu	33,3275	2,360909091	6,216458	2,360909091
8	Lampung	34,1975	3,230909091	7,086458	3,230909091
9	Kep. Bangka Belitung	31,41	0,443409091	4,298958	0,443409091
10	Kep. Riau	30,1825	0,784090909	3,071458	0,784090909
11	DKI Jakarta	27,03	3,936590909	0,081042	0,081041667
12	Jawa Barat	32,6125	1,645909091	5,501458	1,645909091
13	Jawa Tengah	29,8075	1,159090909	2,696458	1,159090909
14	DI Yogyakarta	24,12	6,846590909	2,991042	2,991041667
15	Jawa Timur	27,67	3,296590909	0,558958	0,558958333
16	Banten	31,755	0,788409091	4,643958	0,788409091
17	Bali	25,4325	5,534090909	1,678542	1,678541667
18	Nusa Tenggara Barat	29,78	1,186590909	2,668958	1,186590909
19	Nusa Tenggara Timur	27,9025	3,064090909	0,791458	0,791458333
20	Kalimantan Barat	29,395	1,571590909	2,283958	1,571590909
21	Kalimantan Tengah	29,975	0,991590909	2,863958	0,991590909
22	Kalimantan Selatan	27,1925	3,774090909	0,081458	0,081458333
23	Kalimantan Timur	25,6725	5,294090909	1,438542	1,438541667
24	Kalimantan Utara	28,585	2,381590909	1,473958	1,473958333
25	Sulawesi Utara	29,4075	1,559090909	2,296458	1,559090909
26	Sulawesi Tengah	32,355	1,388409091	5,243958	1,388409091
27	Sulawesi Selatan	27,9075	3,059090909	0,796458	0,796458333
28	Sulawesi Tenggara	28,705	2,261590909	1,593958	1,593958333

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

No	Provinsi	Rata-rata	C1	C2	Jarak Terpendek
29	Gorontalo	32,89	1,923409091	5,778958	1,923409091
30	Sulawesi Barat	29,7	1,266590909	2,588958	1,266590909
31	Maluku	27,435	3,531590909	0,323958	0,323958333
32	Maluku Utara	31,17	0,203409091	4,058958	0,203409091
33	Papua Barat	29,5875	1,379090909	2,476458	1,379090909
34	Papua	27,68	3,286590909	0,568958	0,568958333

Berikut hasil perhitungan pada data klaster literasi 3 yaitu :

Tabel 11. Hasil Klaster Literasi 3

Klaster	Hasil
C1	22
C2	12

Berdasarkan perhitungan manual pada data Persentase merokok yang telah dilakukan diatas mendapatkan hasil akhir literasi 3 bernilai sama yaitu C1= 22 dan C2= 12. Posisi pada literasi tidak berubah, maka proses berhenti dan disimpulkan:

Tabel 12. Provinsi Klaster Tinggi dan Klaster Rendah

No	Provinsi Klaster Tinggi
1	Aceh
2	Sumatera Utara
3	Sumatera Barat
4	Riau
5	Jambi
6	Sumatera Selatan
7	Bengkulu
8	Lampung
9	Kep. Bangka Belitung
10	Kep. Riau
11	Jawa Barat
12	Jawa Tengah
13	Banten
14	Nusa Tenggara Barat
15	Kalimantan Barat
16	Kalimantan Tengah
17	Sulawesi Utara
18	Sulawesi Tengah
19	Gorontalo
20	Sulawesi Barat
21	Maluku Utara
22	Papua Barat

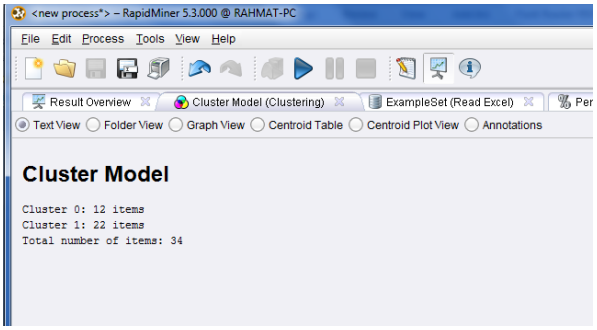
Tabel 13. Provinsi Klaster Rendah

No	Provinsi Klaster Rendah
1	DKI Jakarta
2	DI Yogyakarta
3	Jawa Timur
4	Bali
5	Nusa Tenggara Timur
6	Kalimantan Selatan
7	Kalimantan Timur
8	Kalimantan Utara
9	Sulawesi Selatan
10	Sulawesi Tenggara
11	Maluku
12	Papua

3.3 Hasil Pengujian

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 2, Nomor 2, Januari 2021
e-ISSN 2685-998X
DOI 10.30865/json.v2i2.2770

Setelah mendapatkan hasil akhir dari perhitungan manual, selanjutnya yaitu menyesuaikan hasil dari perhitungan manual menggunakan software Rapidminer 5.3. Disimpulkan bahwa data yang digunakan adalah valid. Hal ini dibuktikan dengan dilakukannya perhitungan manual dan data uji dari *Rapidminer5.3* yang dapat menampilkan hasil yang sama. Tampilan data *klaster model* berupa teks yang ada di *Rapidminer* bisa dilihat pada gambar 1. berikut.



Gambar 1. Klaster Model Data Persentase Merokok Umur Di Atas 15 Tahun

4. KESIMPULAN

Berdasarkan pembahasan sebelumnya dapat disimpulkan Penerapan *Data mining* dengan menggunakan algoritma *K-means* dapat diterapkan. Sumber data yang digunakan pada penelitian ini diambil dari Badan Pusat Statistik (BPS) pada data persentase merokok umur di atas 15 tahun, dari tahun 2015-2018. Jumlah record yang digunakan sebanyak 34 provinsi dengan menghasilkan dua klaster yakni klaster tinggi sebanyak 22 provinsi dan klaster rendah sebanyak 12 provinsi. Berdasarkan hasil pengujian tools *rapidminer* diperoleh hasil yang sama dengan analisis perhitungan *K-means* dan diperoleh 22 provinsi sebagai *klaster* tinggi dimana provinsi tersebut menjadi pusat perhatian lebih kepada pihak pemerintah untuk dapat mengurangi angka persentase merokok di Indonesia.

REFERENCES

[1] S. R. Ningsih, I. S. Damanik, A. P. Windarto, H. S. Tambunan, J. Jalaluddin, and A. Wanto, "Analisis K-Medoids Dalam Pengelompokkan Penduduk Buta Huruf Menurut Provinsi," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 721, 2019.
[2] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*. Deepublish, 2020.
[3] I. W. S. Wicaksana, D. Aprilia, D. A. Baskoro, and Liat Ambarwati, "Belajar Data Mining Dengan Rapid Minner," p. 139, 2013.
[4] K. U. Pengelompokkan, K. Kota, M. W. Talakua, Z. A. Leleury, and A. W. Talluta, "ANALISIS CLUSTER DENGAN MENGGUNAKAN METODE PROVINSI MALUKU BERDASARKAN INDIKATOR INDEKS PEMBANGUNAN MANUSIA TAHUN 2014 CLUSTER ANALYSIS BY USING K-MEANS METHOD FOR GROUPING OF DISTRICT / CITY IN MALUKU PROVINCE INDUSTRIAL BASED ON INDICATORS OF MALUKU DEVELOPMENT INDEX IN 2014," vol. 11, pp. 119–128, 2017.
[5] R. A. Malik, S. Defit, and Y. Yuhandri, "Comparison of K-Means Clustering Algorithm with Fuzzy C-Means In Measuring Satisfaction Level Of Television Da'wah Surau TV," *Rabit*, vol. 3, no. 1, pp. 10–21, 2018.
[6] I. Parlina, A. P. Windarto, A. Wanto, and M. R. Lubis, "MEMANFAATKAN ALGORITMA K-MEANS DALAM MENENTUKAN PEGAWAI YANG LAYAK MENGIKUTI ASESSMENT CENTER UNTUK CLUSTERING PROGRAM SDP," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 3, no. 1, pp. 87–93, 2018.
[7] G. Prasetya, "PENGARUH PENDIDIKAN, PELATIHAN, JENIS KELAMIN, UMUR, STATUS PERKAWINAN, DAN DERAH TEMPAT TINGGAL TERHADAP LAMA MENCARI KERJA TENAGA KERJA TERDIDIK DI INDONESIA," vol. 10, no. 2, pp. 1–15, 2018.