PENERAPAN ALGORITMA K-MEANS UNTUK PENGELOMPOKAN DIAGNOSA PENYAKIT MATA BERDASARKAN RENTANG USIA

Fitra Kurnia[1], Ichsan Fahmi[2], Erwin Wahyudi[3], Godlief E.S. Mige[4]
[1,3]*UIN Sultan Syarif Kasim Fakultas Sains & Teknologi Riau,Indonesia*
Email: fitra.k@uin-suska.ac.id
[2, 4]*Program Studi Pendidikan Teknik Elektro Universitas Nusa Cendana,Kupang Indonesia*
Email : ichsan.fahmi@staf.undana.ac.id

***Abstrak* - Penelitian ini menggunakan teknik Data Mining yang bertujuan untuk menemukan informasi baru dari suatu dataset rekam medis penyakit mata.Algoritma K Means dijadikan pilihan untuk tujuan tersebut. Algoritma ini mengelompokan jenis penyakit mata berdasarkan rentang usiadengan parameter jenis kelamin, umur, gejala dan penyakit.Setelah melalui tahapan seleksi, *preprocessing/cleaning* dan transformasi maka data yang digunakan menjadi hanya 6689 record. Data ini dikelompokkan dalam 3 cluster yaitu penyakit banyak (C1), sedang (C2) dan sedikit (C3). Berdasarkan Algoritma K Means diperoleh informasi diagnosa penyakit terbanyak terjadi pada kelompok Usia Tua yang rentan terhadap penyakit *Catarac*t, kemudian pada kelompok usia balita dan anak-anak, remaja dan dewasa rentan terhadap penyakit *Conjunctivitis.* Tahap pertama, pengujian dilakukan dengan cara membandingkan perhitungan manual dengan aplikasi yang telah dirancang. Hasil perbandingannya menunjukan nilai yang sama. Metode pengujian pertama adalah menggunakan perbandingan *Between-Class Variation* (BCV) dan *Within-Class Variation* (WCV).Rasio perbandingan BCV dan WCV adalah0,002yang bermakna tingkat penggunaan nilai *centroid* memiliki kualitas yang sangat baik.Metode pengujian kedua dilakukan untuk melihat akurasi hasil *clustering*yang dihitung menggunakan Metode *Receiver Operating Characteristic* (ROC).Hasil perhitungan dengan metode ROC adalah0,645.Nilai akurasi ini menunjukan bahwaaplikasi berada dalam kategori baik.Pada tahap kedua, pengujian dilakukan untuk membandingkan hasil aplikasi dengan Rapidminer. Perbandingan pengujian menunjukan bahwa hasil cluster tidak memiliki selisih lebih dari 3% untuk tiap clusternya, dimana cluster 1 memiliki selisih 0 data (0%), cluster 2 memiliki selisih 108 data(2,2%), dan cluster 3 memiliki selisih 108 data(2,2%).**

Keywords: Algoritma K Means,Centroid,Clustering, Data Mining, Preprocessing,Rapidminer

1.PENDAHULUAN

Sebagai salah satu indera yang vital bagi manusia maka mata telah menjadi pusat perhatian dunia.Semua aktifitas yang dilakukan manusia pada dasarnya berasal dari penyerapan informasi visual perlu mendapatkan perhatian serius.

Data gangguan penglihatan di seluruh dunia diperoleh dari hasil estimasi yang dilakukan oleh WHO.Klasifikasi gangguan penglihatan yang digunakan adalah berdasarkan tajam penglihatan.*Low vision* jika tajam penglihatan berkisar <6/18 - ≥3/60 dan buta jika tajam penglihatan kurang dari 3/60.Estimasi jumlah orang dengan gangguan penglihatan di seluruh dunia pada tahun 2010 adalah 285 juta orang atau 4,24% populasi, sebesar 0,58% atau 39 juta orang menderita kebutaan dan 3,65% atau 246 juta orang mengalami low vision. 65% orang dengan gangguan penglihatan dan 82% dari penyandang kebutaan berusia 50 tahun atau lebih[1].

Berdasarkan laporan pada pertemuan Asia Pacific Academy of Ophthalmology di Sydney tahun 2010, Angka kebutaan yang besar di Indonesia menjadi yang tertinggi kedua di dunia setelah Ethiopia, dilaporkan pada pertemuan tersebut bahwa angka kebutaan Indonesia adalah di atas 1%. Angka ini menjadikan kebutaan di Indonesia tidak hanya menjadi masalah kesehatan tetapi sudah menjadi masalah sosial[2].

yang terjadi pada mata. Gangguan yang terjadi pada mata berdampak serius pada hampir semua aspek kehidupan manusia.Upaya mencegah dan menanggulangi gangguan penglihatan dan kebutaan

Katarak atau kekeruhan lensa mata merupakan salah satu penyebab kebutaan terbanyak di Indonesia maupun di dunia. Perkiraan insiden katarak adalah 0,1%/tahun atau setiap tahun di antara 1.000 orang terdapat seorang penderita baru katarak. Penduduk Indonesia juga memiliki kecenderungan menderita katarak 15 tahun lebih cepat dibandingkan penduduk di daerah subtropics[2]. Menurut Persatuan Dokter Spesialis Mata[2],estimasi kemampuan operasi katarak oleh dokter-dokter mata di Indonesia pertahunnya berkisar 150.000-180.000. Perhitungan kasar ini menunjukkan bahwa untuk mencapai angka *Cataract Surgical Rate*(CSR) 2000 saja, Indonesia mempunyai *backlog* operasi katarak sebesar 320.000-350.000 per tahunnya. Jumlah ini akan meningkat sesuai dengan meningkatnya jumlah penduduk dan meningkatnya umur harapan hidup mengingat penderita katarak sebagian besar terjadi pada umur >50 tahun.

Sekitar 80% gangguan penglihatan dan kebutaan di dunia dapat dicegah. Dua penyebab terbanyak adalah gangguan refraksi dan katarak, yang keduanya dapat ditangani dengan hasil yang baik dan cost-effective di berbagai negara termasuk Indonesia.Sebagai titik awal perencanaan program penanggulangan kebutaan dan gangguan penglihatan yang direkomendasikan oleh WHO melalui Vision 2020 adalah ketersediaan data mengenai keadaan kebutaan dan gangguan penglihatan di suatu wilayah. Ketersediaan data dan informasi ini sangat penting agar program penanganan kebutaan dan gangguan penglihatan dirancang berdasarkan permasalahan yang muncul di masyarakat sehingga dapat dilakukan perencanaan program yang efektif dan efisien.

Salah satu upaya untuk menyediakan informasi yang penting ditempuh dengan teknik *Data Mining*(DM). Teknik ini digunakan untuk menemukan informasi baru dari tumpukan data.*Clustering* adalah salah satu metode yang penting dalam DM. Algoritma *Clustering* bekerja dengan mengelompokkan obyek-obyek data (pola, entitas, kejadian,unit,hasil observasi) ke dalam sejumlah cluster tertentu[3].K-means merupakan salah satu algoritma clustering[4]. K-Means membagi data kedalam sejumlah kelompok sehingga data yang berkarakteristik sama dimasukkan ke dalam satu kelompok sementara data yang berkarakteristik berbeda dimasukkan dalam kelompok yang lain. Tujuan dari clustering adalah meminimalkan fungsi obyektif yang diset dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi didalam suatu kelompok dan memaksimalkan variasi antar kelompok.*Clustering* merupakan teknik pengelompokan *record* data pada kriteria tertentu, hasil *clustering* diberikan kepada pengguna akhir untuk memberikan gambaran tentang apa yang terjadi pada basis data[5]. *Clustering* melakukan pengelompokkan data tanpa berdasarkan kelas data tertentu. Bahkan *clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. *Clustering* sering digolongkan sebagai metode *unsupervised learning*[6].Salah satu metode *clustering* yang tergolong *unsupervised learning* adalah K-Means.

1. Data Mining

Data Mining merupakan salah satu proses eksplorasi dan analisis data yang memiliki banyak metode dengan kegunaan masing-masing. Data Mining merupakan gabungan dari berbagai bidang ilmu, antara lain basis data, *information retrieval*, statistika, *machine learning* dan sebagainya. Data Mining dapat diterapkan di berbagai bidang, seperti bisnis, kesehatan, asuransi, pemasaran dan perbankan.

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. Kumpulan proses tersebut meliputi: pembersihan data (*data cleaning*), integrasi data (*data integration*), pemilihan data (*data selection*), Transformasi data (*data transformation*),evaluasi pola (*pattern evaluation*), *knowledge presentation.*

Data Mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data *warehouse*, atau penyimpanan informasi lainnya. Data Mining berkaitan dengan bidang ilmu-ilmu lain seperti *database system*, *data warehouse*, statistik, *machine learning*, *information retrival*, dan komputasi tingkat tinggi. Selain itu, Data Mining didukung oleh ilmu lain seperti *neural network*, pengenalan pola, *spatial data analysis*, *image database*, *signal processing*[5]

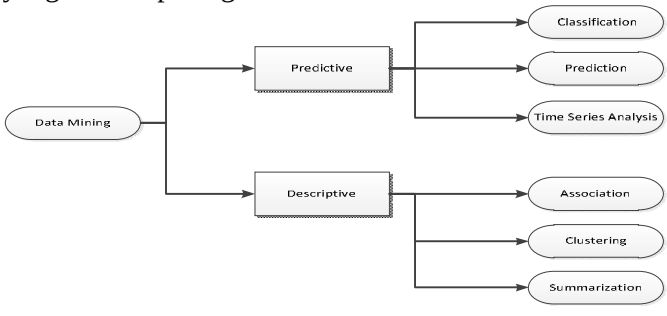
1.1 Karakteristik Data Mining

Data Miningmemiliki beberapa karakteristik tertentu. Berikut adalah tiga karakteristik dari Data Mining[7]:

1) Data Mining berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2) Data Mining bisa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil yang lebih dipercaya.
3) Data Mining hanya berguna untuk membuat keputusan kritis, terutama dalam strategi.

1.2 Tugas Data Mining

Menurut[8]Tugas Data Miningsecara garis besar dibagi menjadi dua kategori utama, seperti yang terlihat pada gambar berikut:



Gambar 1.Tugas Data Mining

Inti dari Data Mining adalah menggali data untuk mendapatkan informasi berharga yang tersembunyi dalam data tersebut. Data Mining

mendukung *task*atau fungsionalitas yang meliputi[8]:

1.Tugas Prediktif

Tujuan ini adalah memprediksi nilai dari atribut tertentu berdasarkan nilai dari atribut lainnya. Atribut yang diprediksi dikenal sebagai target atau *dependent variable*, sedangkan atribut yang digunakan untuk membuat prediksi disebut penjelas atau *independent variable*.Metode yang termasuk prediktive Data Mining:

a) Klasifikasi: pembagian data ke dalam beberapa kelompok atau kelas yang telah ditentukan sebelumnya.
b) Regresi: memetakan data ke suatu *prediction variable*.
c) *Time Series Analysis*: pengamatan perubahan nilai atribut dari waktu ke waktu.

2.TugasDeskriptif

Tujuan utama dari tugas ini adalah untuk memperoleh pola (*correlation, trend, cluster, trajectory, anomaly*) untuk menyimpulkan hubungan di dalam data.Tugas deskriptif merupakan tugas Data Miningyang sering dibutuhkan padateknik *postprocessing* untuk melakukan validasi dan menjelaskan hasil prosesData Mining.Inti dari tugas Data Miningadalah pemodelan prediktif, analisa asosiasi, analisa *cluster*, dan deteksi terhadap anomali.

Metode yang termasuk deskriptive Data Mining:

a. *Clustering*: mengelompokkan beberapa objek yang serupa ke dalam sebuah cluster, dan yang tidak serupa ke cluster yang lain.
b. *Association rules*: identifikasi hubungan antara data yang satu dengan yang lainnya.
c. *Summarization*: pemetaan data ke dalam subset dengan deskripsi sederhana.
d. *Sequence discovery*: identifikasi pola sekuensial dalam data.

Pemodelan prediktif mengacu pada proses membangun model untuk *variable* target sebagai fungsi dari variabel penjelas. Ada dua tipe dari pemodelan prediktif, yaitu klasifikasi (*classification*) yang digunakan untuk variabel target yang diskret, dan regresi (*regression*) yang digunakan untuk variabel target yang kontinyu.Analisa asosiasi digunakan untuk menemukan pola yang mendeskripsikan fitur-fitur data yang saling berhubungan.Pola-pola ini biasanya digambarkan dalam bentuk aturan implikasi. Analisa *cluster* merupakan proses untuk mencari kelompok-kelompok data, sedemikian sehingga data yang berada dalam satu kelompok memiliki kemiripan dibandingkan data yang terletak pada kelompok lain. Deteksi anomaly merupakan proses identifikasi data yang memiliki perbedaan karakteristik yang signifikan dengan data yang lain atau yang dikenal dengan istilah *outlier*[8].

3. Pengelompokkan Data Mining

Data Mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan[9], yaitu:

a.Deskripsi *(Description)*

Terkadang penelitidan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

b.Estimasi *(Estimation)*

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi.

c.Prediksi *(Prediction)*

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa datang.

d.Klasifikasi *(Classification)*

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

e.Pengklasteran *(Clustering)*

Pengklusteran merupakan pengelompokan *record*, pengamatanatau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. *Cluster* adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dalam *cluster* lain.

f.Asosiasi *(Assosiation)*

Tugas asosiasi dalam *Data Mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

4. Clustering

Clustering merupakan teknik pengelompokan *record* data pada kriteria tertentu, hasil *clustering* diberikan kepada pengguna akhir untuk memberikan gambaran tentang apa yang terjadi pada basis data[5]. *Clustering* melakukan pengelompokkan data tanpa berdasarkan kelas data tertentu. Bahkan *Clustering* dapat dipakai untuk memberikan label pada kelas data yang belum diketahui. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning* atau bersifat tanpa supervise[6].

Pengelompokan (*clustering*) merupakan teknik yang sudah cukup dikenal dan banyak digunakan untuk mengelompokkan data atau objek ke dalam

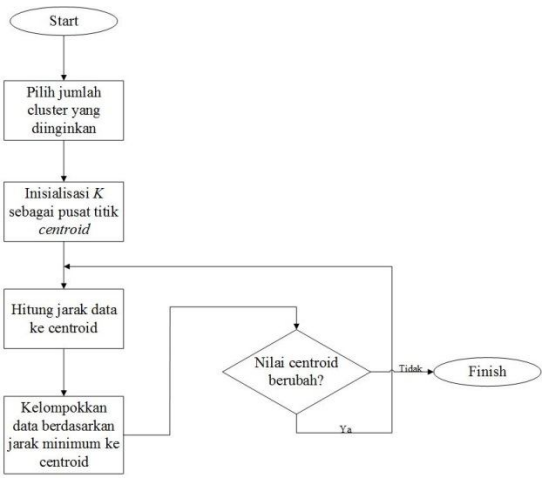
kelompok data (*cluster*) sehingga setiap *cluster* memiliki data yang mirip dan berbeda dengan data yang berada dalam *cluster* lain. Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data/objek ke dalam *cluster* sehingga dalam setiap *cluster* akan berisi data yang sama.

Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu kelas dan meminimumkan kesamaan antar kelas.*Clustering* dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multi dimensi. Jika diberikan himpunan data yang berjumlah terhingga, yaitu *X*, maka permasalahan *clustering* dalam *X* adalah mencari beberapa pusat *cluster* yang dapat memberikan ciri kepada masing-masing *cluster* dalam *X*. [8].

Beberapa penelitian telah dilakukan dengan menggunakan Algoritma K Means dalam beragam kasus.[10] menggunakan K Means untuk clustering data polutan udara di Pekanbaru. Pengujian dilakukan sebanyak 3 kali dengan jumlah cluster yang berbeda -beda.Pengujian pertama dengan jumlah cluster = 5, pengujian kedua dengan jumlah cluster = 4 dan pengujian ketiga dengan jumlah cluster = 3. Hasil penelitian yang penting dari penelitian ini adalah terjadi kenaikan rata – rata volume yang dimulai pada bulan Juni dan kemudian turun kembali pada bulan Otober dan November.Pada kasus yang lain,[11] menggunakan K Means untuk segmentasi pelanggan yang bertujuan untuk mengenali perilaku pelanggan sehingga dapat diterapkan strategi marketing yang tepat.[12] menggunakan K Means untuk clusteringkinerja akademik mahasiswa.Berdasarkan hasil penelitiannya diketahui bahwa pendapatan orang tua tidak mempengaruhi tingkat kinerja akademik mahasiswa dan nilai akademis mahasiswa yang masuk melalui jalur reguler & jalur prestasi akademik mempunyai nilai IPK rata-rata tertinggi. [13],[14],[15] menggunakan K Means untuk meneliti pada subyek yang sama terkait penyakit.[13] fokus pada penyakit menular dengan menggunakan data yang bersumber dari 32 Kantor Puskesmas di Kabupaten Majalengka sedangkan [15] fokus pada penyakit kulit dan kelamin berdasarkan rentang usia menggunakan atribut umur dan diagnosa dengan jumlah data bersih rekam medis = 883.Hasil penelitian [15] dapat disimpulkan bahwa faktor usia dapat mempengaruhi jenis penyakit dari seseorang, seperti hasil pengelompokkan cluster 1 bahwa penyakit infeksi menular seksual 100% diderita oleh pasien berusia dewasa 26 tahun s/d 50 tahun.

2.METODE PENELITIAN

Penelitian ini menggunakan teknik data mining dengan metode Algoritma K Means yang bertujuan mengelompokan data rekam medis penyakit mata.Algoritma ini disajikan dalam flow chart gambar 2.Bagian pertama yang dilakukan adalah menentukan jumlah cluster,kemudian menentukan centroid yang dilakukan secara random pada data yang tersedia.Selanjutnya dilakukan proses hitung jarak data dengan centroid menggunakan Euclidean Distance.Algoritma berakhir jika nilai centroid tidak mengalami perubahan.



Gambar 2. Flow chart Algoritma K Means

1.Algoritma K-Means

K-Means merupakan salah satu metode data *clustering non hirarki* yang berusaha mempartisi data yang ada ke dalam bentuk satu kelompok atau lebih. Metode ini mempartisi data ke dalam *cluster* atau 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 kelompok yang lain[6].

Dalam melakukan *cluster* menggunakan *K-Means*, ada beberapa tahap yang harus dilakukan antra lain :

1. Tentukan jumlah *Cluster* K.
2. Inisialisasikan K titik pusat *cluster* ini dapat dilakukan dengan cara acak dan digunakan sebagai titik pusat *cluster* awal.
3. Alokasikan semua data atau obyek ke *cluster* terdekat. Kedekatan dua obyek ditentukan berdasarkan jarak kedua obyek tersebut. Untuk menghitung jarak semua data kesetiap titik pusat *cluster* menggunakan teori jarak Eucliseaan yang dirumuskan sebagai berikut :

$$(x,y) = \sqrt{(X_{1x} - X_{2y})^2 + (X_{1x} - X_{2y})^2 + \cdots + (X_{1x} + X_{2y})^2} \quad \ldots(1)$$

Dimana :
D (x, y) = Jarak data ke x ke pusat *cluster* y
X kx = Data ke x pada atribut data ke k
Y ky = Titik pusat ke y pada atribut ke k

4. Jarak yang terpendek antara pusat *cluster* dengan data atau obyek menentukan posisi.
5. Hitung kembali pusat *cluster* dengan keanggotaan *cluster* yang baru.

$$C(i) = \frac{x_1 + x_2 + x_3 + x_n}{\sum x} \dots (2)$$

Tugaskan setiap obyek memakai pusat *cluster* yang baru, jika pusat *cluster* berubah kembali kelangkah 3, lakukan proses pengulangan hinggan nilai *cluster* tidak mengalami perubahan.

2.Pengujian Hasil Clustering K-Means

Metode pengujian yang digunakan untuk menentukan kriteria penilaian bagus atau tidaknya hasil dari perhitungan *Clustering K-Means* adalah dengan menggunakan metode *Between-Class Variation* (BCV) dan *Within-Class Variation* (WCV) pada iterasi terakhir yang sering disebut dengan rasio. Apabila hasil perhitungan pengujian yang diperoleh besar, maka semakin bagus tingkat kualitas *clustering* tersebut.

BCV merupakan rata-rata dari *centroid*, sedangkan WCV adalah nilai keseluruhan dari jarak minimum yang telah dijumlahkan. Rumus perhitungannya adalah sebagai berikut :

$$BCV = \frac{1}{Nk} \sum_{i=1}^{k} d(m_i, m_i) \dots\dots\dots\dots (3)$$

Keterangan :
k = Jumlah *cluster*
m_i = Jumlah anggota dari *cluster* ke-*i*
i = Nama yang mewakili *cluster* yang dibentuk
m_i = Jumlah anggota dari *cluster* ke-*i*

$$WCV = \sum_{j=i}^{n} \sum_{p \in ci} d(p, m_i)^2 \dots\dots\dots (4)$$

Keterangan :
p€ci = Jumlah semua data
k = Jumlah *cluster*
p = *Cluster* jarak terdekat
m_i = Jumlah anggota dari *cluster* ke-*i*

$$Rasio = \frac{BCV}{WCV} \dots\dots\dots\dots (5)$$

Apabila nilai rasio yang didapat semakin kecil maka semakin bagus pula tingkat hasil dari akurasi *cluster*, menurut[16].Kriteria hasil ukuran rasio dapat dilihat pada tabel Tabel 1.

Tabel1. Kriteria Pengukuran Rasio

Nilai Rasio	Kriteria
≤ 0,25	Sangat baik
0,25- 0,50	Baik
0,50- 0,75	Kurang baik
0,75– 1,00	Buruk

3. Metode *Receiver Operating Characteristic* (ROC)

Tingkat akurasi diukur dengan cara menggunakan metode ROC.Selain mencari nilai akurasi pada metode ini juga dapat dicari nilai sensitivitas dan spesifitas[17], adapun persamaan nya dapat dilihat sebagai berikut:

$$\text{Akurasi} = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \dots\dots\dots\dots (6)$$

$$\text{Sensifitas} = \frac{Tp}{Tp+Fn} \dots\dots\dots\dots (7)$$

$$\text{Spesifitas} = \frac{Tp}{Tp+Fp} \dots\dots\dots\dots (8)$$

Keterangan :
Tp = True positif (Nilai kebenaran pada nilai *centeroid*)
Tn = True negative (Nilai*centeroid* hasil clustering)
Fp = False positif (Nilai kebeneraan *centeroid* pada cluser lain)
Fn = False Negative (Nilai kebeneraan *centeroid* terakhir pada cluser lain).

Apabila hasil dari *clustering* mendekati titik kurva 1,00 maka akurasi yang didapatkan dalam kategori bagus, untuk melihat hasil akurasi masuk kedalam kategori yang mana, perhatikan tabel Tabel di bawah ini.

Tabel 2.Standar Receiver Operating Characteristic (ROC)

Nilai Rasio	Kategori
0,80 - 1,00	Sangat baik
0,60 - 0,80	Baik
0,40- 0,60	Cukup baik
0,20 - 0,40	Kurang Baik
0,00 - 0,20	Tidak Baik

3.HASIL DAN PEMBAHASAN

Semua algoritma jika digunakan pasti memiliki nilai *error*, semakin kecil nilai *error* yang dimiliki pada suatu sistem maka semakin bagus pula hasil dari kinerja sistem itu. Pada penelitian ini menghitng nilai *error* menggunakan persamaan yang pada sebelumnya sudah dijelaskan, perhitungan nilai error terdapat pada proses berikut ini:

Menentukan iterasi keberapa akan dihitung

Untuk menentukan iterasi diambil pada iterasi terakhir karena iterasi terakhir memiliki kualitas *centeroid* yang lebih baik dari sebelumnya, pada penelitian ini menggunakan sampel data perhitungan penyakit mata, untuk lebih jelasnya perhatikan Tabel di bawah ini.

Tabel 3. Nilai *Centroid* pada iterasi terakhir

	X	Y	Z
C1	1,07	3,19	1,54
C2	2,92	1,38	0,77
C3	4,91	1,33	2,60

Kemudian hitung nilai *Centroid* dengan persamaan (3).

BCV

$$=\sqrt{(1,07-2,92)^2+(3,19-1,38)^2+(1,54-0,77)^2}+\sqrt{(1,07-4,91)^2+(3,19-1,33)^2+(1,54-2,60)^2}+\sqrt{(2,92-4,91)^2+(1,38-1,33)^2+(0,77-2,60)^2}$$

$$= 6,08$$

Menentukan jarak minimum centeroid

Pada proses ini menggunakan jarak minimum pusat *centeroid* yang didapat pada iterasi terakhir, dapat dilihat pada Tabel 4.

Tabel 4.Jarak minimum nilai *Centroid* terakhir

	C1	C2	C3	MIN	CLUSTER
X1	1,76	3,81	1,04	1,04	3
X2	0,71	2,65	2,81	0,71	1
X3	0,84	2,00	2,83	0,84	1
X4	1,70	4,19	0,98	0,98	3
X5	1,74	4,17	1,00	1,00	3
X6	1,70	4,19	0,98	0,98	3
X7	0,84	3,36	1,85	0,84	1
X8	1,52	2,18	3,79	1,52	1
X9	1,70	4,19	0,98	0,98	3
X10	1,70	4,19	0,98	0,98	3
X11	0,71	2,65	2,81	0,71	1
X12	1,76	3,81	1,04	1,04	3
X13	0,84	3,36	1,85	0,84	1
X14	0,92	3,33	1,86	0,92	1
X15	0,92	3,33	1,86	0,92	1
X16	0,92	3,33	1,86	0,92	1
X17	2,50	2,07	4,78	2,07	2
....	.....	.....	.....	.....	.....
X6689	0,84	2,00	2,83	0,84	1

Setelah mendapatkan jarak minimum dengan nilai pusat *centroid* maka langkah selanjutnya hitung seluruh jarak minimum dengan persamaan (4) sebagai berikut:

WCV = $1,04^2+0,71^2+0,84^2+0,98^2+ \ldots+2,05^2+0,84^2$

Sehingga hasil yang didapat adalah WCV = 2653,167

Menghitung perbandingan BCV dengan WCV

Pada langkah terakhir adalah menghitung nilai perbandingan BCV dengan WCV sehingga menghasilkan nilai *error*hitung dengan persamaan (5) seperti terlihat pada hasil dibawah ini.

Rasio = $\frac{6,08}{2653,167}$ = 0,0022916

Untuk menentukan bagus atau tidaknya hasil pengujian dari nilai rasio yang didapat maka harus memperhatikan kriteria pengukuran rasio pada tabel 5.

Tabel 5.Kriteria Pengukuran Rasio

	Kriteria
≤ 0,25	Sangat Baik
0,25 - 0,50	Baik
0,50 - 0,75	Kurang Baik
0,75 – 1,00	Buruk

Hasil pengujian menggunakan perbandingan *Between-Class Variation* (BCV) dan *Within-Class Variation* (WCV) mendapatkan nilai rasio yang tidak tinggi yaitu 0,0022916 dan artinya tingkat penggunaan nilai *Centroid* memiliki kualitas yang sangat baik.

Pengujian Metode ROC

Metode ROC digunakan untuk menghitung nilai akurasi hasil *clustering* yang telah diproses oleh sistem,Selain nilai akurasi, nilai *sensifitas* dan nilai *spesifitas*dapat dihitung juga.Adapun untuk mencari nilai akurasi dapatdicari dengan persamaan (6), untuk mencari nilai *sensifitas* dengan persamaan (7), dan mencari nilai *spesifitas* dengan persamaan (8).Pada penelitian ini digunakan data dari hasil *clustering* rekam medis penyakit mata yang berupa data nilai centeroid awal dan nilai centeroid pada iterasi terakhir.Data tersebut ditampilkan dalam tabel 6 di bawah ini.

Tabel 6.Nilai *Centroid* Data Rekam Medis Penyakit Mata

Kelompok	Centroid Awal	Centroid Iterasi Terakhir
C1	5	1,54
C2	3	1,54
C3	1	2,60

$$\text{Akurasi} = \frac{5+1,54}{5+1,54+1+2,60} = \frac{6,54}{10,14} = 0,645$$

$$\textit{Sensifitas} = \frac{5}{5+1,54} = \frac{5}{6,54} = 0,7645$$

$$\textit{Spesifitas} = \frac{5}{5+8} = \frac{5}{13} = 0,385$$

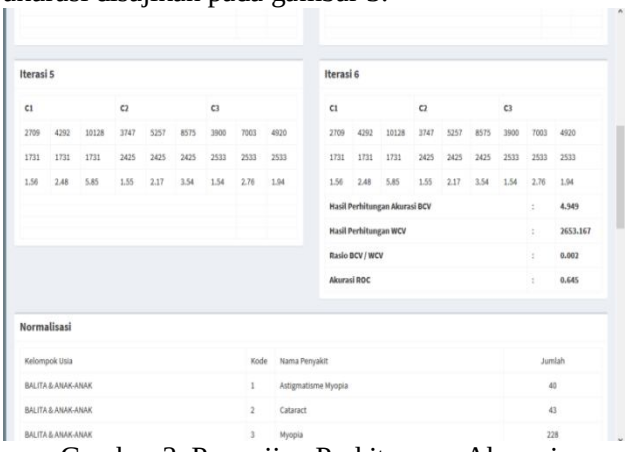
Sesuai dengan perhitungan.nilai akurasi yang didapat adalah0,645. Nilai akurasi ini berada dalam kategori baik berdasarkan referensi pada tabel 7 dibawah ini.

Tabel 7. Nilai Rasio

Nilai Rasio	Kategori
0,80-1,00	Sangat Baik
0,60-0,80	Baik
0,40-0,60	Cukup Baik
0,20-0,40	Kurang Baik
0,00-0,20	Tidak Baik

Pengujian Perhitungan Akurasi Menggunakan Aplikasi

Pengujian perhitungan dilakukan menggunakan Aplikasi yang telah dibangun, dimana terdapat tiga metode perhitungan akurasi yaitu metode ROC, metode BCV dan metode WCV.Hasil perhitungan akurasi disajikan pada gambar 3.

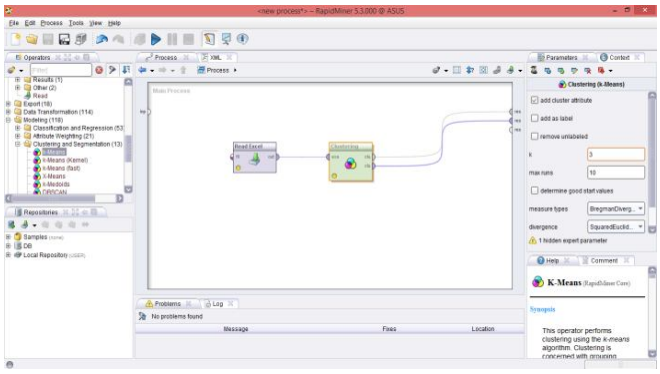


Gambar 3. Pengujian Perhitungan Akurasi Menggunakan Sistem

Pada gambar 3 dapat dijelaskan bahwa perhitungan menggunakan metode BCV dengan WCV menghasilkan nilai rasio sebesar 0,002 sedangkan menggunakan metode ROC menghasilkan nilai akurasi sebesar 0,645.Hal ini menunjukan bahwa perhitungan manual tidak memiliki perbedaan hasil dengan perhitungan yang dilakukan oleh aplikasi.

Pengujian Perbandingan Dengan Tools Rapidminer

Tools yang digunakan dalam pengujian ini adalah Rapidminer Studio versi 5.3.Gambar 4 merupakan gambar design Clustering data penyakit mata menggunakan Rapidminer.Setelah Aplikasi dijalankan, maka akan muncul hasil dari proses perhitungan cluster menggunakan tools Rapidminer seperti yang dapat dilihat pada gambar 5. Setelah tools Rapidminer dijalankan maka. Hasil perhitungan dari Rapidminer selanjutnya dibandingkan dengan hasil perhitungan dari aplikasi.Adapun hasil selisih pengujiannya dapat dilihat pada tabel 8.



Gambar 4. Tampilan Design Tools Rapidminer

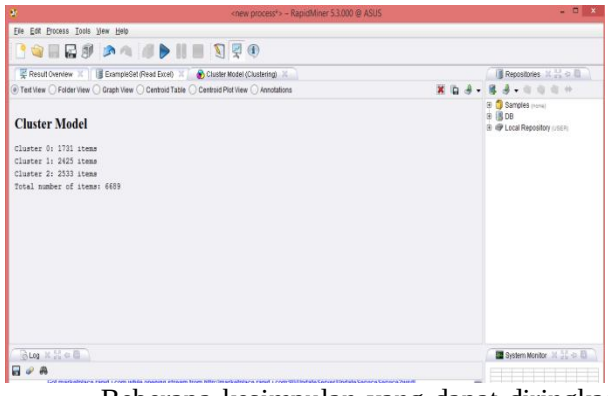
Dari keterangan tabel 8, hasil *Cluster* yang dilakukan secara manual dengan *Rapidminer* tidak memiliki selisih lebih dari 3% dari tiap clusternya,Cluster 1 memiliki selisih 0 data, cluster 2 memiliki selisih 108 data, dan cluster 3 memiliki selisih 108 data.

Tabel 8. Hasil Perbandingan Sistem dan *Tools*

	C1	C2	C3	Total
Sistem	1731	2533	2425	6689
Rapidminer	1731	2425	2533	6689
Selisih	0	108	108	216
Persentase	0%	2,2%	2,2%	3,2%

Gambar 5. Tampilan Hasil Cluster Tools Rapidminer

4.KESIMPULAN



Beberapa kesimpulan yang dapat diringkas dari penelitian ini adalah sebagai berikut :

1. Penyakit mata dikelompokkan menjadi 3 cluster, yaitu kelompok jumlah peyakit Sedang, penyakit Tinggi dan penyakit Rendah sedangkan kategori pasien dikelompokkan menjadi 3, yaitu kelompok usia Balita dan Anak-Anak, usia Remaja dan Dewasa serta kelompok usia Tua.

2. Penyakit *Cataract* menempati urutan pertama sebagai diagnosa penyakit terbanyak berada di kelompok Usia Tua sedangkan pada kelompok usia Balita dan Anak-Anak, Remaja dan Dewasa, rentan terhadap penyakit *Myopia* sebagai diagnosa penyakit terbanyak.
3. Hasil pengujian menggunakan perbandingan *Between-Class Variation* (BCV) dan *Within-Class Variation*(WCV) mendapatkan nilai rasio yang tidak tinggi yaitu 0,0022916 dan artinya tingkat penggunaan nilai centeroid memiliki kualitas yang sangat baik.
4. Berdasarkan akurasi yang didapat pada perhitungan data rekam medis penyakit mata mendapatkan nilai akurasi sebesar 0,645. Dengan nilai akurasi tersebut maka dapat dikategorikan baik.

REFERENSI

[1]World Health Organization,GLOBAL DATA ON VISUAL IMPAIRMENTS,2010

[2]Infodatin,Pusat Data dan Informasi Kementerian Kesehatan RI,2014

[3]Rui Xu,Donald and C. Wunsch, Clustering ,John Wiley & Sons, INC,2009

[4] X. Wu and V. Kumar, eds., The Top Ten Algorithms in Data Mining.Chapman and Hall, 2009

[5]Han, J. & Kamber, M.,2006. Data Mining Concept and Technique. Morgan Kaufman Publisher, San Francisco.

[6] Ian H.Witten, Eibe Frank,2005. Data mining : practical machine learning tools and techniques,Morgan Kaufmann Publisher,San Francisco

[7]Kantardzic, M. 2011. Data Mining: Concepts, Models, Methods and Algorithms.SECOND EDITION, John Wiley & Sons, Inc., Hoboken, New Jersey

[8]Tan, P.N., Steinbach, M. and Kumar, V.2006, Introduction to Data Mining. Pearson Education, Inc., London

[9] Larose, Daniel T.2015, Data mining and predictive analytics,John Wiley & Sons, Inc., Hoboken, New Jersey

[10]Fathuddin Yazid, Muhammad Affandes.Clustering Data Polutan Udara Kota Pekanbaru dengan Menggunakan

Metode K-Means Clustering,Jurnal CoreIT, Vol.3, No.2, Desember 2017,1,Teknik Informatika UIN Sultan Syarif Kasim Riau

[11]Hadi,Fakhri.dkk.2017.Penerapan K-Means Clustering Berdasarkan RFM Mofek Sebagai Pemetaan dan Pendukung Strategi Pengelolaan Pelanggan.Jurnal Sains, Teknologi dan Industri, Vol. 15, No. 1, Desember 2017, pp.69 – 76.Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau

[12]Fajar Nur Rohmat Fauzan Jaya Aziz, Budi Darma Setiawan, Issa Arwani.Implementasi Algoritma K-Means untuk Klasterisasi Kinerja Akademik Mahasiswa.Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X Vol. 2, No. 6, Juni 2018, hlm. 2243-2251.Universitas Brawijaya Malang

[13]Ade Bastian, Harun Sujadi, dan Gigin Febrianto.PENERAPAN ALGORITMA K-MEANS CLUSTERING ANALYSIS PADA PENYAKIT MENULAR MANUSIA.Jurnal Sistem Informasi (Journal of Information System), Volume 14, Issue 1, April 2018. Universitas Majalengka,Jawa Barat

[14]Anindya Khrisna Wardhani.IMPLEMENTASI ALGORITMA K-MEANS UNTUK PENGELOMPOKKAN PENYAKIT PASIEN PADA PUSKESMAS KAJEN PEKALONGAN,JURNAL TRANSFORMATIKA, Volume 14, Nomor 1, Juli 2016,Universitas Diponegoro Semarang

[15]Ni Wayan Wardani dkk,Algoritma K-Means Untuk Pengelompokan Diagnosa Penyakit Kulit Dan Kelamin Berdasarkan Rentang Usia,Prosiding Seminar Nasional Pendidikan Teknik Informatika (SENAPATI 2016),ISSN 2087-2658,Denpasar,Bali

[16]Kauffman,Leonard.Rousseeuw,Peter J 2005.Finding Groups in Data An Introduction to Cluster AnalysisJohn Wiley & Sons, Inc., Hoboken, New Jersey

[17] Vercellis,Carlo 2009,Business Intelligence: Data Mining and Optimization for Decision Making,A John Wiley and Sons, Ltd., Publication,United Kingdom