BAB IV
ANALISA

Analisa adalah tahapan yang akan dikakukan untuk membahas dan memahami masalah yang didapatkan dalam penelitian yang akan dilakukan. Pada tahap analisa ini maka inti dari permasalahan yang ada akan bisa dipahami dengan baik. Data awal untuk penelitian ini yaitu data pasien dari RSUD Rokan Hulu yang cek darah di laboratorium. Dalam tahap analisa ini akan dijelaskan bagaimana data diolah dengan menerapkan metode K-Means sehingga akan dapat kelompok klaster yang diinginkan.

4.1. Analisa kebutuhan data

Data yang digunakan dalam penelitian ini adalah 500 data pasien RSUD Rokanhulu yang cek darah di RSUD ini. Adapun data yang dihasilkan adalah berupa beberapa table yang berisikan informasi tentang umur, gula darah acak (GDA),gula darah puasa (GDP), Trigliserida (TG), kolesterol baik (HDL) dan kolesterol jahat (LDL). Adapun pengertian datanya sebagai berikut :

Tabel 4.1 Atribut data awal

ATRIBUT	KETERANGAN ATRIBUT	TIPE DATA
MR	Medical Record	*Integer*
Umur	Umur pasien	*Integer*
GDA	Gulda Darah Acak Pasien	*Integer*
GDP	Gula Darah Puasa Pasien	*Integer*
TG	Trigliserida Pasien	*Integer*
HDL	Kolesterol baik pasien	*Integer*
LDL	Kolesterol jahat pasien	*Integer*

Berdasarkan Atribut-atribut yang telah dijelaskan pada Tabel 4.1 maka data awal yang akan digunakan dalam penelitian untuk mencari kelompok *cluster* dapat dilihat seperti pada Tabel 4.2.

Tabel 4.2 Data Awal Penelitian

NO	MR	UMUR	GDA	GDP	TG	HDL	LDL
1	270020	67	232	142	122	70	100
2	465780	68	139	93	101	91	82
3	641390	56	210	139	146	54	255
4	641930	57	122	87	90	71	95
5	702570	57	112	78	89	94	115
6	710440	51	256	160	141	42	141
7	719480	63	345	165	223	42	152
8	752200	70	267	200	117	61	225
9	790580	35	100	93	83	81	118
10	791450	69	110	81	89	62	82
…	…	…	…	…	…	…	…
495	795623	43	256	204	122	92	112
496	793426	37	184	132	117	101	134
497	796676	49	291	221	129	125	182
498	795073	54	253	197	126	53	104
499	798481	45	120	94	91	85	81
500	793404	34	245	168	117	62	224

Selain dari data yang akan digunakan dalam penelitian tersebut, pada metode yang akan digunakan yaitu *K-means* juga terdapat data jumlah target *cluster* yang akan ditentukan terlebih dahulu. Target jumlah *cluster* yang digunakan pada penelitian ini terdapat 2 *cluster,* karena pada penelitian ini bermasuksud untuk mengelompokan data pasien kedalam klaster pasien yang terjangkit diabetes mellitus atau tidak terjangkit saja, dapat dilihat pada Tabel 4.3.

Tabel 4.3 Banyak Klaster

Satuan Nilai	Keterangan
1	*Cluster* 1
2	*Cluster* 2

4.2. Proses KDD

Berdasarkan analisa kebutuhan data yang telah dijelaskan di atas maka pada bagian ini dijelaskan tahapan yang akan dilakukan terhadap data pasien pada penelitian ini.

4.2.1. Seleksi Data

Pada tahapan seleksi data dilakukan pemilihan atribut-atribut data yang sesuai dengan kebutuhan. Untuk atirbut yang akan dipakai adalah atribut Umur, GDA, GDP, TG, HDL dan LDL, sedangan untuk atribut MR tidak dipakai karena tidak digunakan dalam perhitungan K-Means. Sehingga atribut yang akan dipakai ada 6 atribut.

4.2.2. *Preprocessing*

Preprocessing merupakan tahap yang akan dilakuan setelah seleksi data pada KDD *(Knowledge Discovery in Database).* Pada tahap ini dilakukan pembersihan data yang mungkin terdapat kosongnya salah satu atribut data atau adanya noise atau kesalahan input data. Akan tetapi pada penelitian ini semua data terisi penuh dan tidak ada salah data sehingga tidak ada yang dihilangkan. Karena tidak adanya data yang kosong atau *missing value* pada semua data pasien, maka tahap ini dilewati dan dilanjutkan ke tahap berikutnya.

4.2.3. Tranformation

Pada tahap ini data yang digunakan untuk penelitian dilakukan tahapan perubahan bentuk atribut dari data yang didapat supaya sesuai dengan proses data mining. Akan tetapi tidak ada data pasien yang atributnya diperlukan perubahan jenis. Contohnya jika ada data yang berbentuk *string* maka dirubah menjadi integer semua.

4.2.4. Data Mining

Data mining merupakan tahap utama pada penelitian ini. Pada penelitian ini digunakan K-Means sebagai algoritmanya. Pada algoritma K-Means akan terbentuk klaster-klaster yang pada setiap klaster memiliki kemiripan karateristik

data dan antara klaster memiliki perbedaan karakterisitik data. Pada penelitian ini perhitungan jarak yang digunakan adalah *euclidean distance* yang merupakan perhitungan jarak dari 2 titik dalam *euclidean space.* Setelah didapatkan jarak maka akan di ambil jarang yang terpendek.

Berikut ini akan dijelaskan langkah tahapan *Data Mining*, data yang digunakan merupakan data hasil inisialisasi menggunakan algoritma *K-Means*.

a. Tentukan banyaknya cluster yang akan dibentuk. Pada penelitian ini *cluster* yang akan dibentuk sebanyak 2 *cluster* karena peneliti ingin mengelompokkan data pasien yang tidak terjangkit diabetes mellitus dan data pasien yang terjangkit dibetes mellitus. Banyaknya *cluster* harus lebih sedikit dari pada banyaknya data (k<n). Data yang bisa digunakan adalah data pasien yang sudah melalui tahapan KDD sebanyak 500 data.
b. Tentukan *centroid* tiap klaster secara random seperti pada tabel berikut ini.

Tabel 4.4 Penentuan centroid tiap klaster

NO	Umur	GDA	GDP	TG	HDL	LDL
1	67	232	142	122	70	100
2	68	139	93	101	91	82
3	56	210	139	146	54	255
4	57	122	87	90	71	95
5	57	112	78	89	94	115
6	51	256	160	141	42	141
7	63	345	165	223	42	152
8	70	267	200	117	61	225
9	35	100	93	83	81	118
10	69	110	81	89	62	82
...	...	...	...	...	...	...
156	67	113	100	96	84	84
157	66	331	166	114	65	150
158	60	131	65	88	67	113
...	...	...	...	...	...	...
499	45	120	94	91	85	81
500	34	245	168	117	62	224

Untuk pengulangan pada perhitungan berikutnya, *centroid* baru dihitung dengan menghitung nilai rata-rata data pada tiap *cluster*. Jika *centroid* baru berbeda

dengan *centroid* sebelumnya maka proses dilanjutkan. Namun jika *centroid* baru sama dengan *centroid* sebelumnya maka proses *Clustering* selesai.

c. Hitung jarak data dengan *centroid* menggunakan *Euclidean Distance* dengan rumus sebagai berikut:

$$D(Xj - Cj) = \sqrt{\sum_{j=0}^{n}(Xj - Cj)^2}$$

Keterangan

D = Jarak

X = Data

C = *Cluster/Centroid*

j = Data

Tabel 4.5 Centroid awal

No	Umur	GDA	GDP	TG	HDL	LDL	
No. 5	57	112	78	89	94	115	C1
No. 157	66	331	166	114	65	150	C2

$$\boldsymbol{D(X1, Y1) = \sqrt{(Ai - Aj)^2 + (Bi - Bj)^2 + (Ci - Cj)^2 + (Di - Dj)^2 + (Ei - Ej)^2}}$$

1. Jarak antara data pertama dengan centroid pertama:

$d_{1,0}\sqrt{(67-57)^2 + (232-112)^2 + (142-78)^2 + (122-89)^2 + (70-94)^2 + (100-115)^2}$ = 143,129

2. Jaraka antara data pertama dengan centroid kedua:

$d_{1,1}\sqrt{(67-66)^2 + (232-331)^2 + (142-166)^2 + (122-114)^2 + (70-65)^2 + (100-150)^2}$ =113,873

3. Jarak antara data kedua dengan centroid pertama:

$d_{2,0}\sqrt{(68-57)^2 + (139-112)^2 + (93-78)^2 + (101-89)^2 + (91-94)^2 + (82-115)^2}$=48,135

4. Jarak antara data kedua dengan centroid kedua:

$d_{2,1}\sqrt{(68-66)^2 + (139-331)^2 + (93-166)^2 + (101-114)^2 + (91-65)^2 + (82-150)^2}$=218,325

5. Jarak antara data ketiga dengan centroid pertama:

$d_{3,0}\sqrt{(56-57)^2 + (210-112)^2 + (139-78)^2 + (146-89)^2 + (54-94)^2 + (255-115)^2}$=194,358

6. Jarak antara data ketiga dengan centroid kedua:

$d_{3,1}\sqrt{(56-66)^2 + (210-331)^2 + (139-166)^2 + (146-114)^2 + (54-65)^2 + (255-150)^2}$=166,253

7. Jarak antara data keempat dengan centroid pertama:

$d_{4,0}\sqrt{(57-57)^2 + (122-112)^2 + (87-78)^2 + (90-89)^2 + (71-94)^2 + (95-115)^2}$=33,33

8. Jarak antar data keempat dengan centroid kedua:

$d_{4,1}\sqrt{(57-66)^2+(122-331)^2+(87-166)^2+(90-114)^2+(71-65)^2+(95-150)^2}$231,603

9. Jarak antara data kelima dengan centroid pertama:

$d_{5,0}\sqrt{(57-57)^2+(112-112)^2+(78-78)^2+(89-89)^2+(99-94)^2+(115-115)^2}$=0

10. Jarak antara data kelima dengan centroid kedua:

$d_{5,1}\sqrt{(57-66)^2+(112-331)^2+(78-166)^2+(89-114)^2+(99-65)^2+(115-150)^2}$=241,82

11. Jarak antara data keenam dengan centroid pertama:

$d_{6,0}\sqrt{(51-57)^2+(256-112)^2+(160-78)^2+(141-89)^2+(42-94)^2+(141-115)^2}$=183,248

12. Jarak antara data keenam dengan centroid kedua:

$d_{6,1}\sqrt{(51-66)^2+(256-331)^2+(160-166)^2+(141-114)^2+(42-65)^2+(141-150)^2}$=85

13. Jarak antara data ketujuh dengan centroid pertama:

$d_{7,0}\sqrt{(63-57)^2+(345-112)^2+(165-78)^2+(223-89)^2+(42-94)^2+(152-115)^2}$=289,695

14. Jarak antara data ketujuh dengan centroid kedua:

$d_{7,1}\sqrt{(63-66)^2+(345-331)^2+(165-166)^2+(223-114)^2+(42-65)^2+(152-150)^2}$=112,339

15. Jarak antara data kedelapan dengan centroid pertama:

$d_{8,0}\sqrt{(70-57)^2+(267-112)^2+(200-78)^2+(117-89)^2+(61-94)^2+(225-115)^2}$=230,328

16. Jarak antara data kedelapan dengan centroid kedua:

$d_{8,1}\sqrt{(70-66)^2+(267-331)^2+(200-166)^2+(117-114)^2+(61-65)^2+(225-150)^2}$=104,489

17. Jarak antara data kesembilan dengan centroid pertama:

$d_{9,0}\sqrt{(35-57)^2+(100-112)^2+(93-78)^2+(83-89)^2+(81-94)^2+(118-115)^2}$=32,665

18. Jarak antara data kesembilan dengan centroid kedua:

$d_{9,1}\sqrt{(35-66)^2+(100-331)^2+(93-166)^2+(83-114)^2+(81-65)^2+(118-150)^2}$=248,781

19. Jarak antara data kedua puluh dengan centroid pertama:

$d_{10,0}\sqrt{(69-57)^2+(110-112)^2+(81-78)^2+(89-89)^2+(62-94)^2+(82-115)^2}$=47,64

20. Jarak antara data kedua puluh dengan centroid kedua:

$d_{10,1}\sqrt{(69-66)^2+(110-331)^2+(81-166)^2+(89-114)^2+(62-65)^2+(82-150)^2}$=247,665

Dilanjutkan dengan menghitung jarak antara data ke *centroid 1* hingga *centroid* 2 hingga selesai. Hasil dari perhitungan jarak antara data dengan ke 2 *centroid* dapat dilihat seperti pada Tabel 4.6 di bawah ini.

Tabel 4.6 Jara antara data dengan centroid awal

No	C1	C2	terpendek
1	143.1293	113.8727	113.8727
2	48.13523	218.3254	48.13523
3	194.3579	166.2528	166.2528
4	33.33167	231.6031	33.33167
5	0	241.8202	0
6	183.2485	85	85
7	289.6947	112.3388	112.3388
8	230.328	104.4892	104.4892
9	32.66497	248.781	32.66497
10	47.64452	247.655	47.64452
...	...	...	...
495	194.7049	99.17157	99.17157
496	98.45811	158.6411	98.45811
497	244.1393	98.80789	98.80789
498	192.93	97.94386	97.94386
499	41.29165	236.2964	41.29165
500	200.0175	117.9746	117.9746

d. Kelompokkan data berdasarkan *clusternya*

Kelompokkan data sesuai dengan *cluster* nya, yaitu data yang memiliki jarak terpendek. Contoh pada data pertama dari Tabel 4.7 dapat di lihat bahwa jarak data ke *centroid* 1 lebih kecil daripada centroid yang lainnya. Kelompokkan semua data berdasarkan jarak terpendek seperti contoh data pertama, hasil dari pengelompokkan jarak terdekat bisa dilihat seperti tabel 4.7 di bawah ini.

Tabel 4.7 Kelompok jarak terdekat antara data dan centroid

NO	HASIL
1	C2
2	C1
3	C2
4	C1
5	C1
…	…
496	C1
497	C2
498	C2
499	C1
500	C2

e. Proses kembali lagi ke langkah nomor 2 yaitu dengan menggunakan *centroid* baru dari iterasi pertama yang dihitung dari nilai rata-rata tiap kelompok *cluster*. Untuk *Centroid* baru didapatkan dari jumlah seluruh dari sebuah atribut pada satu *Centroid* dibagi dengan jumlah data, dan begitu seterusnya untuk semua atribut *Centroid.* Sebagai contoh untuk atribut umur pada *Centroid* pertama:

$$\frac{\text{jumlah seluruh umur pada klaster pertama}}{\text{banyak data pada klaster pertama}}$$

Tabel 4.8 Data klaster pertama

No	Data C1					
	Umur	GDA	GDP	TG	HDL	LDL
1	68	139	93	101	91	82
2	57	122	87	90	71	95
3	57	112	78	89	94	115
…	…	…	…	…	…	…
258	37	184	132	117	101	134
259	45	120	94	91	85	81
Total	12599					

$$\frac{12599}{259} = 48.6$$

Contohnya untuk atribut *Centroid* baru untuk pengulangan pertama dapat dilihat pada Tabel 4.9 dibawah ini

Tabel 4.9 Centroid baru untuk iterasi 2

	Umur	GDA	GDP	TG	HDL	LDL
C1	48.625	124.0795	85.23485	96.97727	79.25758	101.7462
C2	50.919492	273.2797	181.0636	149.6059	58.16102	191.5085

Proses dilanjutkan dengan menggunakan *centroid* baru dilakukan pengulangan secara terus menerus selama nilai rata-rata tiap *cluster* berubah dan kemudian berhenti pada pengulangan ke 4 dimana *centroid* terakhir yang tidak mengalami perubahan dapat dilihat pada Tabel 4.9 dibawah ini.

Tabel 4.10 Centroid terakhir yang tidak mengalami perubahan

	Umur	GDA	GDP	TG	HDL	LDL
C1	48.644788	122.8378	84.0695	96.2973	79.7722	100.1583
C2	50.850622	271.5187	180.3278	149.2448	58.04564	191.3527

Tabel persentasi yang didapatkan dari proses *Data Mining* dengan menggunakan algoritma *K-Means* beserta nama untuk masing-masing *cluster* . untuk klaster pertama didapatkan hasil sebanyak 259 data dari seluruh data yang berjumlah 500 data. Sehingga persentasi klaster pertama yaitu 52% data dan pada klaster kedua 48 % data karena memiliki 241 data dari 500 data yang tersedia. Hasil ini juga dapat dilihat pada lampiran B4 dan dapat dilihat pada tabel 4.10 dibawah ini.

Tabel 4.11 Hasil *Clustering*

CLUSTER 1 (C1)	**259**	52 %	Normal
CLUSTER 2 (C2)	**241**	48 %	Diabetes mellitus
TOTAL	**500**		