BAB II

LANDASAN TEORI

2.1 Diabetes

Diabetes adalah penyakit kronis yang terjadi baik ketika pankreas tidak menghasilkan cukup insulin atau ketika tubuh tidak dapat secara efektif menggunakan insulin yang dihasilkan. Insulin adalah hormon yang mengatur gula darah. Hiperglikemia, atau timbul gula darah, adalah efek umum dari diabetes yang tidak terkontrol dan dari waktu ke waktu menyebabkan kerusakan serius pada banyak sistem tubuh, terutama saraf dan pembuluh darah dan ada 2 jenis diabetes yaitu (WHO,2016) :

- Diabetes tipe 1

  Diabetes tipe 1 (sebelumnya dikenal sebagai insulin-dependent, remaja atau awal masa kanak-kanak) ditandai dengan produksi insulin kekurangan dan membutuhkan pemberian sehari-hari untuk insulin 3. Penyebab diabetes tipe 1 tidak diketahui dan tidak dapat dicegah dengan pengetahuan saat ini. Gejala termasuk ekskresi berlebihan urin (poliuria), rasa haus (polidipsia), kelaparan konstan, penurunan berat badan, perubahan pengelihatan dan kelelahan. Gejala-gejala ini dapat terjadi tiba-tiba.

- Diabetes tipe 2

  Diabetes tipe 2 (sebelumnya disebut non-insulin-dependent atau orang dewasa-onset) hasil dari penggunaan yang tidak efektif tubuh insulin 3. Diabetes tipe 2 terdiri mayoritas penderita diabetes di seluruh dunia, dan sebagian besar merupakan hasil dari kelebihan berat badan dan kurangnya

aktivitas fisik. Gejala mungkin mirip dengan diabetes tipe 1, tetapi sering kurang ditandai. Akibatnya, penyakit ini dapat didiagnosis beberapa tahun setelah penyerangan, setelah komplikasi telah muncul. Sampai saat ini, diabetes tipe ini terlihat hanya pada orang dewasa tetapi kini juga terjadi semakin sering pada anak-anak (WHO, 2016).

2.2 Data Mining

Menurut Ginting (2014), *data mining* adalah kegiatan yang meliputi pengumpulan, pemakaian data historis yang menemukan keteraturan, pola dan hubungan dalam set data berukuran besar. Karakteristik *data mining* sebagai berikut :

1. *Data mining* berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2. *Data mining* biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih terpecaya.
3. *Data mining* berguna untuk membuat keputusan yang kritis, terutama dalam strategi.

Secara umum ada dua jenis metode dalama *data mining,* metode *predictive* dan metode *descripstive*. Metode *predictive* adalah proses untuk menemukan pola dari data yang menggunakan beberapa variable untuk memprediksi variable lain yang tidak diketahui jenis atau nilainya. Teknik yng terasuk dalam *predictive mining* antara lain Klasifikasi, Regresi dan Deviasi. Metode *descriptive* adalah proses untuk menemukan suatu karakteristik penting dari data dalam suatu basis data. Teknik *data mining* yang termasuk dalam *descriptive mining* adalah *clustering, association* dan *sequential mining*.

Tahapan *data mining* menurut Ridwan (2013) ada beberapa jenis, yaitu

1. Pembersihan data (*data cleaning*)

   Pembersihan data merupakan proses menghilangkan *noise* dan data yang tidak konsisten atau tidak relevan.

2. Integrasi data (*data integration*)

   Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.

3. Seleksi data (*data selection*)

   Data yang dari database seringkali tidak semuanya akan digunakan dalam melakukan *data mining,* oleh karena itu hanya data yang sesuai kebutuhan akan dipakai dan dianalisis yang akan diambil dari *database*.

4. Transformasi data (*data transformation*)

   Data diubah atau digabungkan ke dalam format yang sesuai untuk diproses dalam *data mining*.

5. Proses *mining*

   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Terdapat beberapa metode yang dapat digunakan dalam pengelompokkan *data mining* seperti *naïve bayes, K-means, Neural Network, C4.5* dll. Pada penelitian kali ini yang digunakan adalah C4.5.

6. Evaluasi Pola (*pattern evaluation*)

   Untuk mengidentifikasi pola-pola menarik ke dalam *knowledge based* yang ditemukan.

7. Presentasi pengetahuan (*knowledge presentation*)

Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh oleh pengguna.

2.3 Pohon Keputusan

Menurut Slamat (2013), merupakan salah satu metode yang ada pada teknik klasifikasi dalam data mining. Metode pohon keputusan mengubah data yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Pohon keputusan juga berguna untuk mengekplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan suatu parameter yang disebut sebagai kriteria dalam pembentukan pohon. Misalkan untuk menentukan main tenis, kriteria yang diperhatikan adalah cuaca, angin, dan suhu. Salah satu atribut merupakan atribut yang menyatakan data solusi per item data yang disebut atribut hasil. Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, C4.5, CART.

2.4 Algoritma C4.5

Algoritma C4.5 yaitu sebuah algoritma yang digunakan untuk membangun decision tree (pengambilan keputusan) (Hanik, 2011). Algortima C4.5 merupakan salah satu algortima *machine learning*. Dengan algoritma ini, mesin (komputer) akan diberikan sekelompok data untuk dipelajari yang disebut *learning dataset*. Kemudian hasil dari pembelajaran selanjutnya akan digunakan untuk mengolah data-data yang baru yang disebut *test dataset*. Karena algoritma C4.5 digunakan untuk melakukan klasifikasi, jadi hasil dari pengolahan *test*

dataset berupa pengelompokkan data ke kelas-kelasnya (Rudi Bening, Tanpa Tahun).

Menurut Andriani (2013), ada beberapa tahapan dalam membuat sebuah *decision tree* dalam algorita C4.5, yaitu:

1. Mempersiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
2. Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, hitung dahulu nilai entropy. Untuk menghitung nilai entropy digunakan rumus :

$$\text{Entropy (S)} = \sum_{j=1}^{k} - Pj\ Log_2\ Pj \quad \ldots (2.1)$$

Keterangan :

Entropy : pengukuran yang berdasarkan kemungkinan yang digunakan untung menghitung jumlah ketidaktentuan.

S : Himpunan kasus

N : jumlah partisi S

P_j : Proporsi S_j terhadap S

Kemudian hitung nilai *gain* menggunakan rumus :

$$\text{Gain (A)} = \text{Entropi (S)} - \sum_{i=1}^{n} \frac{|Si|}{|S|} \times Entropi\ (Si) \quad \ldots (2.2)$$

Keterangan:

Gain : Salah satu *attribute selection measure* yang digunakan untuk memilih *test attribute* tiap *node* pada tree.

A : Atribut

n : Jumlah partisi atribut A

$|S_i|$: Jumlah kasus pada partisi ke-i

$|S|$: Jumlah kasus dalam S

3. Ulangi langkah ke-2 dan langkah ke-3 hingga semua record terpartisi.
4. Proses partisi pohon keputusan akan berhenti saat:
   a. Semua *record* dalam simpul N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam record yang dipartisi lagi.
   c. Tidak ada *record* di dalam cabang yang kosong.

No	Cuaca	Suhu	Kelembaban	Berangin	Main
1	Cerah	Panas	Tinggi	Salah	Tidak
2	Cerah	Panas	Tinggi	Benar	Tidak
3	Berawan	Panas	Tinggi	Salah	Ya
4	Hujan	Sejuk	Tinggi	Salah	Ya
5	Hujan	Dingin	Normal	Salah	Ya
6	Hujan	Dingin	Normal	Benar	Ya
7	Berawan	Dingin	Normal	Benar	Ya
8	Cerah	Sejuk	Tinggi	Salah	Tidak
9	Cerah	Dingin	Normal	Salah	Ya
10	Hujan	Sejuk	Normal	Salah	Ya
11	Cerah	Sejuk	Normal	Benar	Ya
12	Berawan	Sejuk	Tinggi	Benar	Ya
13	Berawan	Panas	Normal	Salah	Ya
14	Hujan	Sejuk	Tinggi	Benar	Tidak

Gambar 2.1 Contoh tabel hitung C4.5 (Rudi Bening, Tanpa Tahun).

Berdasarkan Gambar 2.1 berisikan isi dari contoh tabel yang akan dicari nilainya akan bermain atau tidak. Terdapat 5 atribut dari tabel tersebut yaitu cuaca, suhu kelembaban, berangin dan main. Yang akan menjadi index dari atribut tersebut adalah tabel main karena nanti akan ditentukan hasilnya adalah bermain atau tidak. Pertama kali yang akan di hitung adalah nilai dari *entropy* keseluruhan data yang ada pada setiap atribut.

Contoh penghitungan *entropy* total:

(S) = (-(10/14) x $\log_2$(10/14) + (-4/14) x $\log_2$(4/14)) = 0.863120569

Atribut	Nilai	Sum(Nilai)	Ya	Tidak	Entropy	Gain
Cuaca	Berawan	4	4	0	0	
	Hujan	5	4	1	0.721928095	
	Cerah	5	2	3	0.970950594	
						0.258521037
Suhu	Dingin	4	4	0	0	
	Panas	4	2	2	2	
	Sejuk	6	4	2	2	
						0.183850925
Kelembaban	Tinggi	7	3	4	0.985228136	
	Normal	7	7	0	0	
						0.370506501
Berangin	Salah	8	6	2	0.811278124	
	Benar	6	2	4	0.918295834	
						0.005977711

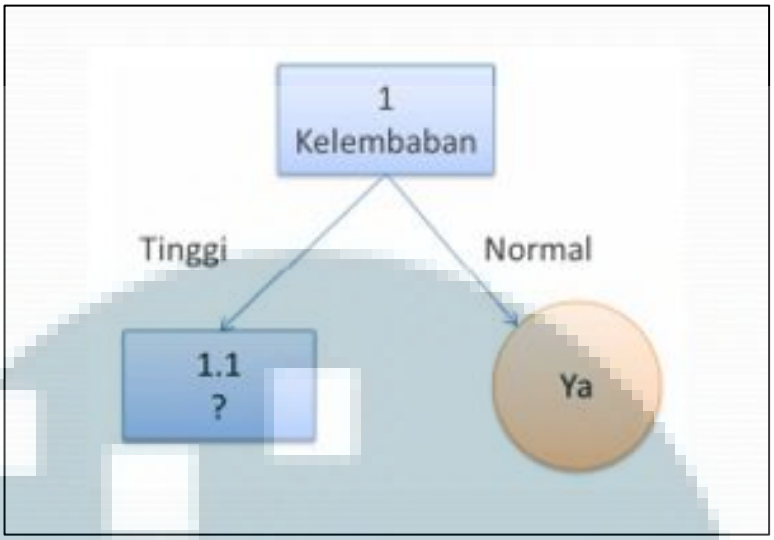
Gambar 2.2 Contoh penghitungan *entropy* dan *Gain* (Rudi Bening, Tanpa Tahun).

Gambar 2.2 menunjukkan penghitungan entropy dan gain dari setiap elemen dari atribut tanpa atribut main karena hasil akhir yang akan dicari adalah main. Kemudian dari hasil penghitungan *entropy* dan *gain* dari setiap elemen dari atribut akan di cari *node* pertama dengan melihat nilai *gain* tertinggi.

Contoh penghitungan *Gain:*

Gain (cuaca) = 0.863120569(entroypy total) –((4/14) x 0 + (5/14) x 0.721928095 + (5/14) x 0.970950594 = 0.258521037

Nilai *Gain* tertinggi yang didapatkan dari Gambar 2.2 adalah kelembaban dengan nilai 0.370506501 maka kelembaban akan menjadi *node* pertama pada *tree* yang akan dibentuk.



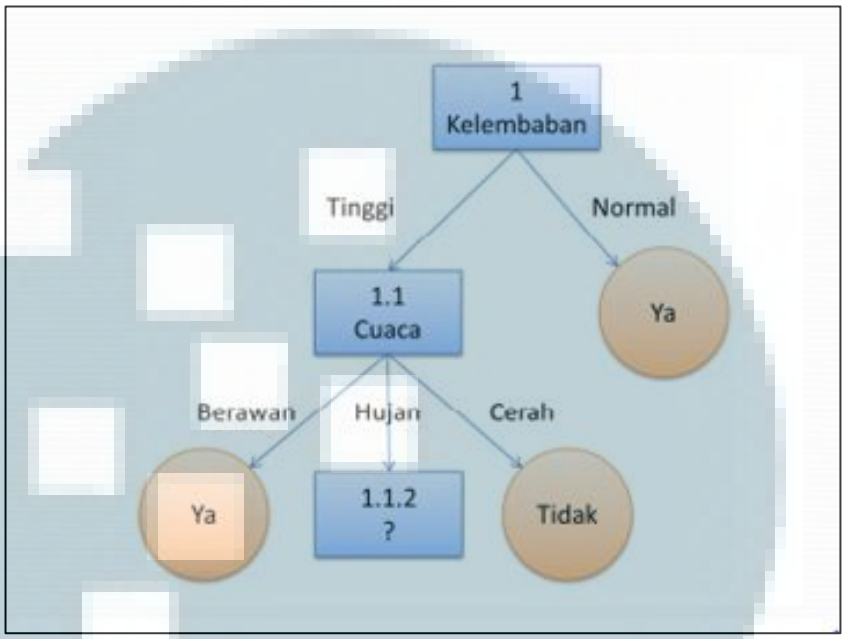
Gambar 2.3 Contoh *tree* yang terbentuk (Rudi Bening, Tanpa Tahun).

Tree yang terbentuk dari hasil penghitungan akan membentuk seperti gambar 2.3. Setelah mendapatkan *node* pertama dari *tree* kemudian dicari lagi atribut kedua yang akan menjadi cabang dari kelembaban dengan menghitung nilai *entropy* dan gain tanpa *node* yang telah dipilih menjadi *tree* dengan sesuai dengan kriteria dari cabang *node* pertama yaitu tinggi.

Atribut	Nilai	Sum(Nilai)	Ya	Tidak	Entropy	Gain
Cuaca	Berawan	2	2	0	0	
	Hujan	2	1	1	1	
	Cerah	3	3	0	0	
						0.69951385
Suhu	Dingin	0	0	0	0	
	Panas	3	2	1	0.918295834	
	Sejuk	4	2	2	1	
						0.020244207
Berangin	Salah	4	2	2	1	
	Benar	3	2	1	0.918295834	
						0.020244207

Gambar 2.4 Sisa tabel atribut setelah penghitungan pertama (Rudi Bening, Tanpa Tahun).

Dari Gambar 2.4 didapatkan nilai *gain* tertinggi adalah cuaca, maka cuaca dijadikan sebagai *node* kedua setelah kelembaban dan akan membentuk tree dibawah tinggi.



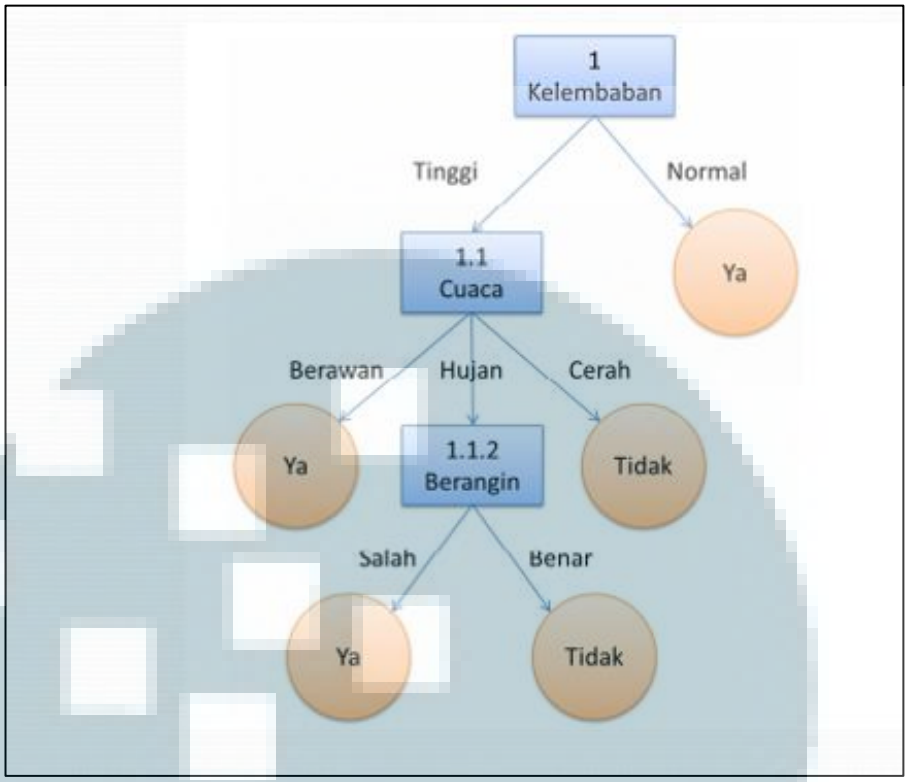
Gambar 2.5 Pembentukkan *tree* lanjutan (Rudi Bening, Tanpa Tahun).

Kemudian hitung kembali lagi nilai *gain* dan *entropy* pada atribut yang tersisa selain yang telah menjadi *tree* dan sesuai dengan spesifik dari cabang *tree* cuaca yaitu hujan.

Atribut	Nilai	Sum(Nilai)	Ya	Tidak	Entropy	Gain
Suhu	Dingin	0	0	0	0	
	Panas	0	0	0	0	
	Sejuk	2	1	1	1	
						0
Berangin	Salah	1	1	0	0	
	Benar	1	0	1	0	
						1

Gambar 2.6 Sisa tabel atibut dari penghitungan kedua (Rudi Bening, Tanpa Tahun).

Nilai *gain* tertinggi yang didapatkan adalah berangin yang menjadi *node* selanjutnya pada pembentukkan *tree*.



Gambar 2.7 hasil akhir *tree* yang terbentuk (Rudi Bening, Tanpa Tahun).

Karena sudah tidak ada cabang lagi yang bisa dicari, maka *tree* selesai terbentuk seperti Gambar 2.7 dengan kelembaban sebagai *node* pertama, kemudian cuaca *node* kedua dan berangin *node* terakhir dalam pembentukkan *tree*.

2.5 Cross-Validation

Menurut Rafaelizadeh (2008), Cross Validation adalah metode statistik yang digunakan untuk mengevaluasi dan membandingkan algoritima pembelajaran dengan membagi data menjadi dua buah segmen: satu digunakan sebagai data *training* dan satu lagi sebagai *testing*. Salah satu teknik dari cross validation adalah *k-fold Cross Validation*, validasi yang dilakukan adalah dengan membagi seluruh data sampel menjadi k-*segment* secara merata, jumlah *segment* disesuaikan dengan jumlah uji coba, sebanyak k. Setiap uji coba yang dilakukan, setiap segmen ke-*k* akan dipilih menjadi data *testing* dan *segment*

lainnya akan digunakan sebagai data *training*. Uji coba akan dilakukan sebanyak *k* dan semua *segment* telah selesai diuji sebagai data *testing*.

2.6 Readmission

Readmission dalam rumah sakit adalah kejadian ketika seseorang yang telah keluar dari rumah sakit tersebut kemudian kembali lagi dalam jangka waktu tertentu, janga waktu yang biasanya digunakan adalah 30 hari setelah keluar dari rumah sakit, sama seperti yang digunakan oleh *Center for Medicare and Medicaid Services (CMS).* Mengurangi readmission adalah salah satu hal penting untuk mengetahui potensi dalam meningkatkan kualitas dalam pelayanan rumah sakit dan mengurangi pengeluaran dalam pengobatan. (American Hospital Asociation,2015)

Terjadinya *readmission* yang dalam jangka pendek dapat menunjukkan kualitas pelayanan di rumah sakit, mencerminkan buruknya koordinasi pelayanan ketika pasien di rawat dan perencanaan atau persiapan kepulangan dari rumah sakit yang tidak lengkap. (The National Center for Biotechnology, 2016)