BAB III
LANDASAN TEORI

3.1 Data Mining

Data mining adalah serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data (Ian H. Witten, 2011). Data mining akan memecahkan masalah dengan menganalisis data yang telah ada dalam basis data. Data mining, sering juga disebut *Knowledge Discovery in Database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan pola keteraturan, pola hubungan dalam set data berukuran besar (Santoso, 2007). Hasil keluaran dari data mining ini dapat dijadikan untuk memperbaiki pengambilan keputusan di masa depan. Pekerjaan (*task*) yang berkaitan dengan dengan data mining dapat dibagi menjadi empat kelompok : model prediksi (*prediction modelling*), analisis cluster (*cluster analysis*), analisis asosiasi (*association analysis*), dan deteksi anomali (*anomaly detection*).

Model prediksi berkaitan dengan pembuatan sebuah model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian model tersebut digunakan untuk memberikan nilai target pada himpunan baru yang didapat. Ada 2 jenis model prediksi, yaitu klasifikasi dan regresi. Klasifikasi digunakan

3.2 Metode Klasifikasi Data Mining

Klasifikasi adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Kamber, 2006). Klasifikasi data terdiri dari 2 langkah proses. Pertama adalah *learning* (fase *training*), dimana algoritma klasifikasi dibuat untuk menganalisis data *training* lalu direpresentasikan dalam bentuk rule klasifikasi. Proses kedua adalah klasifikasi, dimana data tes digunakan untuk memperkirakan akurasi dari rule klasifikasi (Kamber, 2006). Proses klasifikasi didasarkan pada empat komponen (Gorunescu, 2011): a. kelas, variabel dependen yang berupa

kategorikal yang merepresentasikan ‘label’ yang terdapat pada objek. Contohnya: resiko penyakit jantung, resiko kredit, customer loyalty, jenis gempa. b. *predictor*, variabel independen yang direpresentasikan oleh karakteristik (atribut) data. contohnya: merokok, minum alkohol, tekanan darah, tabungan, aset, gaji c. *training dataset* satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan kelas yang cocok berdasarkan predictor d. *testing dataset*, berisi data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi.

3.3 K-Nearest Neighbor

K-Nearest Neighbor (KNN) merupakan salah satu metode yang digunakan dalam menyelesaikan masalah pengklasifikasian. Prinsip KNN yaitu mengelompokkan atau mengklasifikasikan suatu data baru yang belum diketahui kelasnya berdasarkan jarak data baru itu ke beberapa tetangga (*neighbor*) terdekat. Tetangga terdekat adalah objek latih yang memiliki nilai kemiripan terbesar atau ketidakmiripan terkecil dari data lama. Jumlah tetangga terdekat dinyatakan dengan k. Nilai k yang terbaik tergantung pada data. Pendekatan sederhana untuk menentukan nilai k yaitu : k= √n , dimana n adalah jumlah dari sampel data yang ada. Misalkan terdapat 30 sampel data, untuk menentukan nilai k nya digunakan rumus √n √30 5,47 , berarti nilai k = 5 √n √20 4,47 , berarti nilai k = 4.

Nilai k umumnya ditentukan dalam jumlah ganjil (3, 5, 7) untuk menghindari munculnya jumlah jarak yang sama dalam proses pengklasifikasian. Apabila terjadi dua atau lebih jumlah kelas yang muncul sama maka nilai k menjadi k – 1 (satu tetangga kurang), jika masih ada yang sama lagi maka nilai k menjadi k – 2 , begitu seterusnya sampai tidak ditemukan lagi kelas yang sama banyak. Banyaknya kelas yang paling banyak dengan jarak terdekat akan menjadi kelas dimana data yang dievaluasi berada. Dekat atau jauhnya tetangga (neighbor) biasanya dihitung berdasarkan jarak Euclidean (Euclidean Distance). Berikut rumus pencarian jarak menggunakan rumus Euclidian (1):

$$d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \quad (1)$$

dengan:

x1 = sampel data
x2 = data uji
i = variabel data
dist = jarak
p = dimensi data

3.4 Gizi

Pengertian gizi adalah segala asupan yang diperlukan agar tubuh menjadi sehat. Gizi diperoleh dari asupan makanan yang mengandung karbohidrat, protein, lemak, vitamin, dan mineral. Ada tiga macam kondisi dalam penilaian status gizi, yaitu : 1. ditujukan untuk perorangan atau untuk kelompok masyarakat. 2. pelaksanaan pengukuran satu kali atau berulang secara berkala. 3. situasi dan kondisi pengukuran baik perorangan atau kelompok masyarakat pada saat kritis, darurat, kronis, dan sebagainya. Dengan memperhatikan ketiga macam kondisi tersebut, beberapa penilaian status gizi dapat diaplikasikan, seperti penapisan (*screening*), penilaian status gizi perorangan untuk keperluan rujukan dari kelompok masyarakat atau dari puskesmas, dalam kaitannya dengan tindakan atau intervensi. Selain itu dapat dimanfaatkan untuk penilaian status gizi pada kelompok masyarakat dalam rangka mengevaluasi suatu program atau sebagai bahan perencanaan atau penetapan kebijakan. Ada berbagai cara untuk menilai status gizi, salah satunya adalah pengukuran tubuh manusia yang dikenal dengan istilah Antropometri.

3.5 IMT (Indeks Massa Tubuh)

IMT bisa memperkirakan lemak tubuh, tetapi tidak dapat diartikan sebagai persentase yang pasti dari lemak tubuh. Hubungan antara lemak dan IMT dipengaruhi oleh usia dan jenis kelamin. Wanita lebih mungkin memiliki persentase lemak tubuh yang lebih tinggi dibandingkan pria dengan nilai BMI

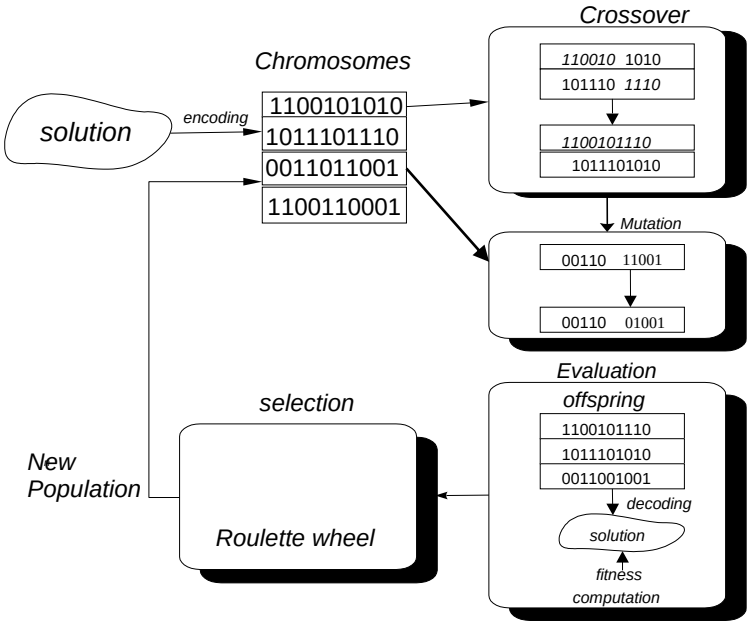
yang sama. Pada BMI yang sama, orang yang lebih tua memiliki lebih banyak lemak tubuh dibandingkan orang yang lebih muda. Rumus BMI adalah sebagai berikut seperti pada (2) :

$$BMI = \frac{berat(kg)}{(tinggi(m))^2} \quad (2)$$

Perhitungan menggunakan rumus BMI menghasilkan kriteria sebagai berikut:
Kurangdari 18,5 : Kurus
18,5 - 22,9 : Normal
Lebih dari 22,9 : digolongkan Obesitas.
Kriteria tersebut merupakan kriteria BMI untuk ukuran orang Asia.

3.6 Algoritma Genetika

Algoritma genetika adalah algoritma yang dikembangkan dari proses pencarian solusi menggunakan pencarian secara acak dan berdasarkan proses seleksi alamiah dan genetik (Gen dan Cheng, 1997). Secara umum Gen dan Cheng (1997) menggambarkan struktur umum algoritma genetika seperti pada Gambar 3.9.



Gambar 3.9 Struktur umum algoritma genetika (Gen dan Cheng, 1997)

Berdasarkan Gambar 3.9, algoritma genetika dimulai dari sekumpulan solusi acak yang disebut populasi awal. Masing-masing individu yang ada dalam populasi dinyatakan sebagai kromosom yang merepresentasikan suatu solusi. Kromosom-kromosom tersebut akan mengalami iterasi yang disebut generasi. Pada setiap generasi, kromosom-kromosom tersebut dievaluasi berdasarkan nilai *fitness*-nya. Untuk menghasilkan kromosom dan generasi baru yang disebut *offspring*, dilakukan penggabungan dua kromosom dari generasi tersebut dengan menggunakan operator *crossover* dan dengan memodifikasi kromosom dengan operator mutasi. Generasi baru dibentuk melalui proses seleksi berdasarkan nilai *fitness*-nya. Populasi pada setiap generasi akan terus dijaga berada pada jumlah yang tetap.

3.6.1 Individu

Menurut Sivanandam dan Deepa (2008) setiap invidu merupakan satu calon solusi. Setiap individu berkelompok membentuk dua bentuk solusi yaitu,

1. Kromosom, yang merupakan informasi dasar genetika (*genotype*)
2. *Phenotype,* yang merupakan model dari sebuah kromosom.

Sebuah kromosom terbagi menjadi gen-gen. Sebuah gen merupakan representasi dari sebuah calon solusi. Masing-masing solusi berkorespondensi dengan gen di dalam kromosom. Satu solusi diwakili dengan satu kromosom yang berbeda satu dengan yang lainnya.

3.6.2 Gen

Gen merupakan instruksi dasar untuk membangun sebuah algoritma genetika. Sebuah kromosom merupakan rangkaian dari gen. Gen dapat dapat mendeskripsikan solusi dari permasalahan tetapi bukanlah solusi itu sendiri. Sebuah gen terdiri dari bit string yang direpresentasikan secara biner dengan interval dari batas bawah ke batas atas. Struktur masing-masing gen ditunjukkan dalam sebuah record yang memetakan genotype dengan phenotype. Pemetaan tersebut penting untuk merubah set solusi dari model menjadi bentuk yang dapat digunakan oleh operator algoritma genetika (Sivanandam dan Deepa, 2008).

3.6.3 Fitness

Sivanandam dan Deepa (2008) menyatakan bahwa fitness individu dalam algoritma genetika merupakan nilai dari fungsi obyektif *phenotype*nya. Kromosom harus dikodekan dan fungsi obyektifnya harus dievaluasi terlebih dahulu sebelum dilakukan perhitungan fitness. Fitness tidak hanya mengindikasikan solusi yang terbaik, tetapi juga memperlihatkan kedekatan kromosom dengan titik optimalnya.

3.6.4 Populasi

Sebuah populasi merupakan kumpulan individu. Sebuah populasi terdiri dari sejumlah individu yang diuji dan ditentukan oleh parameter *phenotype* ditambah informasi lainnya yang ada dalam proses pencarian. Dua aspek penting populasi yang digunakan dalam algoritma genetika adalah :

1. Populasi generasi awal
2. Ukuran populasi

Ukuran populasi akan bergantung pada kerumitan permasalahan dan dalam proses inisialiasi ukuran populasi secara acak atau dipastikan sebelumnya. Secara ideal, populasi pertama harus memiliki sebuah kumpulan gen sebesar mungkin sehingga dapat mengeksplorasi keseluruhan kandidat solusi atau mengarah pada solusi. Semua kemungkinan solusi harus dimunculkan dalam populasi, yang dalam hal ini dilakukan secara acak (Sivanandam dan Deepa, 2008).

3.6.5 Komponen-komponen Algoritma Genetika

Dalam pencarian solusi optimal menggunakan algoritma genetika, pada dasarnya terdapat enam komponen. Tetapi banyak metode yang bervariasi pada masing-masing komponen, sehingga akan ditetapkan metode yang digunakan pada masing-masing komponen yang akan diimplementasikan.

a. Skema pengkodean

Pengkodean kromosom yang merupakan representasi set parameter, merupakan hal yang utama dalam memanipulasi kromosom. Terdapat dua skema pengkodean kromosom, yaitu *real-number encoding* dan *binary encoding* (Sivanandam dan Deepa, 2008).

Pada kromosom yang terdapat 6 parameter yang berisi bilangan yang bernilai real yang dikodekan ke dalam sebuah kromosom yang terdiri dari 6 gen kemudian gen-gen pada kromosom pada saat inisialisasi populasi dibangkitkan secara acak. Masing-masing kromosom akan dikodekan menjadi individu dengan nilai *fitness* tertentu berdasarakan fungsi *fitness*.

b. Fungsi fitness

Sivanandam dan Deepa (2008) mennyatakan bahwa kromosom atau individu dievaluasi berdasarkan fungsi tertentu sebagai ukuran performansinya. Di dalam evolusi alam, individu yang bernilai tinggi yang akan dapat bertahan hidup. Sedangkan individu dengan nilai *fitness* rendah akan mati.

Fungsi *fitness* ini kemudian diuraikan dengan metode linear fitness ranking yang kemudian diseleksi menggunakan metode seleksi *roulette wheel*.

c. Seleksi orang tua

Seleksi orang tua digunakan untuk memilih kromosom-kromosom bernilai *fitness* tinggi yang akan diberikan kesempatan reproduksi yang lebih besar. Nilai tersebut tergantung dari nilai acak yang dibangkitkan. Pemilihan kromosom-kromosom dilakukan dengan menggunakan seleksi *roulette wheel*. Penggunaan metode seleksi ini karena metode ini mudah diimplementasikan dan penerapannya yang sederhana.

d. Pindah silang

Pindah silang digunakan untuk mendapatkan kromosom dengan solusi yang baik atau diharapkan mendapat nilai *fitness* yang tinggi. Proses pindah silang dilakukan dengan memindahsilangkan dua buah kromosom.

3.6.6 Penyandian Kromosom

Algoritma genetik mempunyai kemampuan untuk mencetak inisialisasi populasi pada wilayah *feasible solution*, dan merekombinasi populasi-populasi dengan harapan bahwa pencarian solusi adalah yang benar-benar menjanjikan pada wilayah *state space.* Setiap *feasible solution* di-encoding sebagai suatu kromosom, dan setiap kromosom diberikan nilai *fitness*-nya. Ukuran nilai *fitness* tersebut yang menunjukkan kemampuan untuk bertahan hidup dan menghasilkan keturunan (Gen dan Cheng, 1997).

Parameter dikodekan dalam rentangan dengan nilai *integer* yang tersusun linear membentuk suatu individu. Masing-masing bit, disebut gen, membawa informasi tentang sifat dari individu.

3.7.7 Operasi Reproduksi

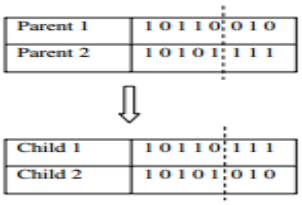
Reproduksi adalah proses pembangkitan individu baru, yang meliputi:

a. *Crossover* (persilangan)

Crossover adalah proses persilangan dari dua individu (yang dianggap sebagai induk) untuk menghasilkan individu baru yang memiliki sifat dari kedua induknya (Sivanandam dan Deepa, 2008). Macam-macam *crossover* adalah sebagai berikut :

- *One point*

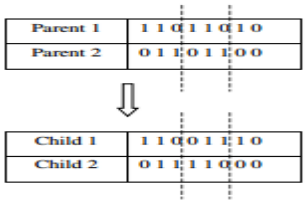
Kromosom pada dua induk ditukarkan di titik *crossover* yang diseleksi secara *random*, dengan sisi yang satu tetap dan yang lain ditukarkan dengan kedua kromosom pada titik *crossover* seperti pada Gambar 3.10.



Gambar 3.10 *One point crossover* (Sivanandam dan Deepa, 2008)

- *Two point*

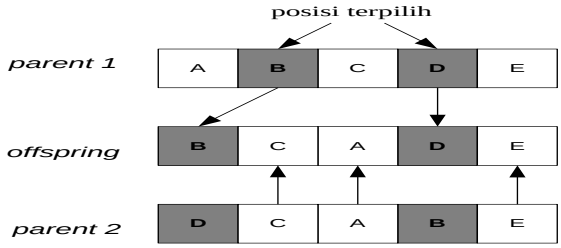
Dua kromosom dipotong di dua titik untuk dipertukarkan bagian tengahnya, seperti pada Gambar 3.11.



Gambar 3.11 *Two point crossover* (Sivanandam dan Deepa, 2008)

- *Order based crossover*

Crossover diawali dengan memilih gen dari kromosom yang dijadikan *parent* 1. Kemudian *parent* 2 yang gen-nya bukan posisi terpilih pada *parent 1* di-*copy*-kan ke turunannya *(offspring)* menurut urutan posisi yang bersesuaian dengan *parent* 2. Selanjutnya posisi terpilih *parent 1*ditempatkan ke *offspring* yang belum terisi dari kiri ke kanan menurut urutannya. Hapus nilai yang sudah ada di *offspring* dari *parent* 2, sehingga didapat *offspring* hasil *crossover.* Ilustrasi *Order based crossover* ditunjukkan pada Gambar 3.12.



Gambar 3.12 *Order Based Crossover* lima gen (Gen Cheng, 1997)

- *Uniform*

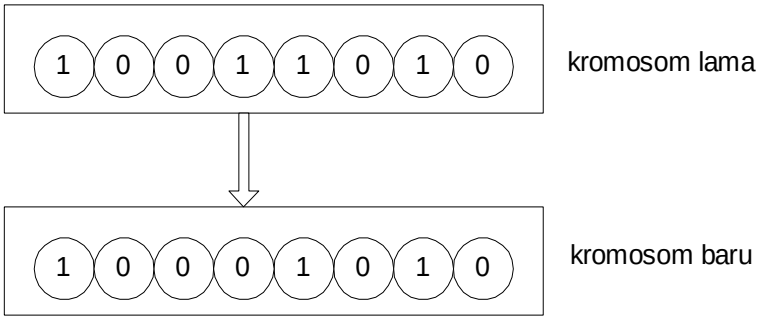
Beberapa gen pada dua kromosom induk saling dipertukarkan secara acak. Ilustrasi *uniform crossover* ditunjukkan pada Gambar 3.13.

Parent 1	1 0 1 1 0 0 1 1
Parent 2	0 0 0 1 1 0 1 0
Mask	1 1 0 1 0 1 1 0
Child 1	1 0 0 1 1 0 1 0
Child 2	0 0 1 1 0 0 1 1

Gambar 3.13 *Uniform crossover* (Sivanandam dan Deepa , 2008)

b. Mutasi

Kemampuan mewarisi sifat induknya (perubahan dalam urutan DNA) disebut mutasi. Mutasi dalam algoritma genetik, biasanya menggunakan pola *gene inversion* (pembalikan gen). Gen atau bit yang dikehendaki mutasi dibalik (dari 0 menjadi 1 dan dari 1 menjadi 0). Kondisi ini untuk menjaga agar tidak ada informasi pada kromosom yang hilang. Dengan mutasi ini setiap kromosom baru dapat diciptakan dengan melakukan modifikasi terhadap satu atau lebih gen pada kromosom yang sama, seperti ditunjukkan Gambar 3.14.



Gambar 3.14 *Gen/bit inversion* gen ke empat (Gen dan Cheng,1997)

c. *Elitism*

Suatu operasi penggantian individu terburuk dalam generasi yang baru dengan individu yang terbaik dari generasi induk. Penggantian dilakukan dengan lebih dahulu menyeleksi individu terbaik dalam generasi induk. Individu dengan *fitness* paling tinggi pada sebuah generasi akan dipertahankan untuk masuk ke generasi berikutnya. Tanpa *elitism* ini ada kemungkinan individu terbaik dalam suatu generasi akan hilang oleh operasi-operasi, mutasi , *crossover* atau seleksi.

3.6.8 Kontrol Parameter Genetik

Pemilihan parameter genetik menentukan kinerja algoritma genetik dalam memecahkan suatu masalah. Parameter ini terdiri atas ukuran populasi, *crossover rate,* dan *mutation rate* (Sivanandam dan Deepa, 2008).

- **Ukuran populasi**

Ukuran ini menjelaskan berapa banyak individu dalam suatu generasi. Jika jumlah individu terlalu sedikit, maka ruang pencarian solusi menjadi sempit/terbatas. Sebaliknya jika jumlah individu tiap generasi terlalu banyak, akan diperlukan waktu yang lama untuk mengevaluasi seluruh individu.

- ***Crossover rate*** **(laju persilangan)**

Crossover rate menunjukkan probabilitas operasi *crossover* dua individu. Semakin besar *crossover rate* semakin cepat struktur individu baru diperkenalkan dalam populasi.

- ***Mutation rate*** **(laju mutasi)**

Mutation rate menunjukkan probabilitas perubahan secara acak suatu gen dalam kromosom, untuk menghasilkan individu baru. Semakin besar laju mutasi semakin cepat lahirnya gen baru pada suatu individu.