BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1. Tinjauan Pustaka

Teknik klasifikasi merupakan pendekatan untuk menjalankan fungsi klasifikasi dalam Data Mining yaitu untuk menggolongkan data. Teknik klasifikasi ini dapat pula digunakan untuk melakukan prediksi atas informasi yang belum diketahui sebelumnya. Beberapa algoritma yang dapat digunakan antara lain adalah algoritma Decision Tree C.45 (Anyanwu dan Shiva, 2009; Hall, 2000; Todorovski dan Džeroski, 2000), algoritma C.50 (Kurgan dan Cios, 2004), *Artificial Neural Networks* (ANN), *K-Nearest Neighbor* (KNN), algoritma *Naive Bayes*, serta algoritma lainnya (Gharehchopogh dkk., 2015).

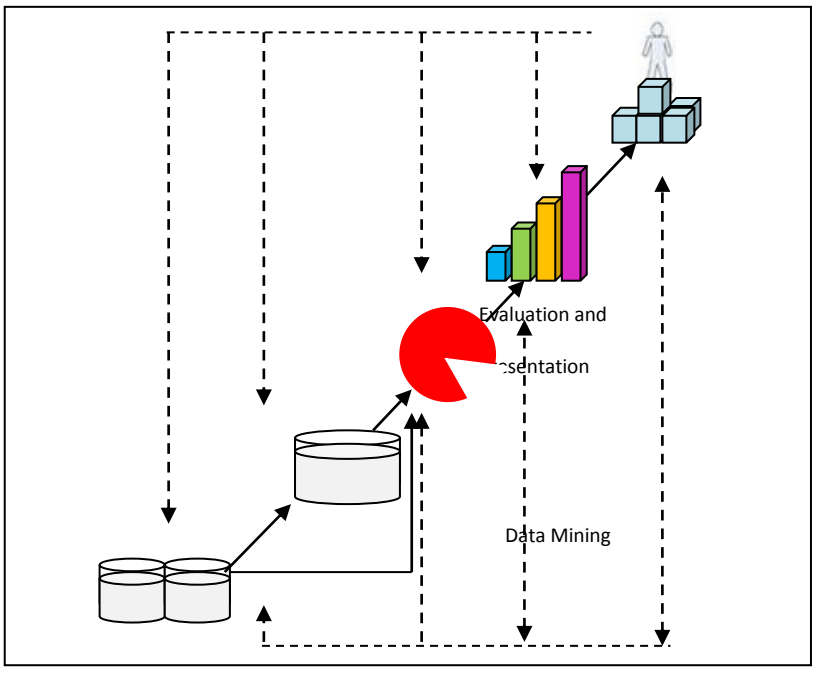
Beberapa penelitian yang menggunakan teknik data mining pada data set Akademik dan Kemahasiswaan telah banyak dilakukan, antara lain adalah menganalisa dan mengevaluasi data akademik dengan menggunakan metode pohon keputusan *(decision tree)* untuk mendapatkan kinerja dari siswa yang selanjutnya dapat digunakan untuk mengetahui kualitas perguruan tinggi (Radaideh dkk., 2006); melakukan penelitian tentang klasifikasi mahasiswa baru berdasarkan prediksi Indeks Prestasi Semester (Studi kasus Program Studi Teknik Informatika Universitas Bina Darma Palembang) dengan menggunakan metode *Case Base Reasoning* (*CBR*) (Pramudyo, 2008), melakukan penelitian untuk mengklasifikasi kinerja akademik mahasiswa dengan menggunakan algoritma *Supervised Learning In Quest* (SLIQ) (Jananto, 2010); dalam penelitian Aplikasi Mining Data Mahasiswa dengan menggunakan metode klasifikasi *Decision Tree* (Sunjaya, 2010); serta penelitian relevan lainnya yang telah membandingkan beberapa algoritma klasifikasi data mining, seperti penelitian yang mengkomparasi algoritma C.45, algoritma *Naive Bayes* dan *Neural Network (*Leidiyana, 2011). Hasil penelitian yang diperoleh dari hasil pengujian dengan mengukur kinerja ketiga algoritma tersebut diketahui bahwa algoritma C.45 memiliki nilai akurasi paling tinggi, diikuti oleh *Neural Network* dan yang terendah adalah *Naive Bayes*.

jaringan saraf memiliki keuntungan relatif sederhana, non-linear, sama dengan informasi fuzzy, persyaratan minimum untuk sumber data, kemampuan untuk belajar contoh-contoh spesifik. Selama proses pembelajaran masukan jaringan syaraf menerima urutan parameter sumber bersama-sama dengan diagnosis yang berasal dari parameter (Yurii dan Liudmila, 2017).

2.2. Landasan Teori

2.2.1. Data Mining

Alasan utama mengapa data mining diperlukan adalah karena adanya sejumlah besar data yang dapat digunakan untuk menghasilkan informasi dan *knowledge* yang berguna. Informasi dan *knowledge* yang didapat tersebut dapat digunakan pada banyak bidang, mulai manajemen bisnis, kontrol produksi, kesehatan, dan lain-lain. Langkah proses datamining dapat dilihat pada Gambar 2.1. (Han dan Kamber, 2001).



Gambar 2.1 Data Mining adalah suatu langkah di dalam proses *Knowledge discovery from data* (KDD) (Han dan Kamber, 2001)

Data mining merupakan suatu rangkaian proses yang dibagi menjadi beberapa tahap seperti yang diilustrasikan di Gambar 2.1 Tahap-tahapnya sebagai berikut:

1. Data *Preprocessing*

   Pada bagian ini menyajikan gambaran dari data *preprocessing*. Pada bagian data quality, mengilustrasikan banyak unsur yang menentukan kualitas data. Ini memberikan insentif balik bagi Data *preprocessing* dan selanjutnya menguraikan tugas utama dalam data *preprocessing*.

2. Pembersihan data (*Data Cleaning*)

   Pembersihan data dilakukan untuk membuang data yang tidak konsisten atau data yang tidak dibutuhkan.

3. Integrasi data (*Data Integration*)

   Integrasi data merupakan penggabungan data dari berbagai sumber. Kekinian populer di industri informasi adalah untuk melakukan integrasi data dan data pembersihan sebagai langkah awal, dimana data yang dihasilkan disimpan dalam *warehouse*.

4. Data *Reduction*

   Data Reduction berguna untuk mendapatkan pengurangan representasi dari kumpulan data yang jauh lebih kecil di dalam volume tetapi belum menghasilkan hasil yang sama (atau hampir sama) dari suatu hasil analisis.

5. *Data Transformation and Data Discretization*

   Dalam *Data Transformation* dan *Data Discretization*, data diubah atau dikonsolidasikan, sehingga proses *mining* yang dihasilkan mungkin lebih efisien, dan pola yang ditemukan mungkin lebih mudah untuk dipahami.

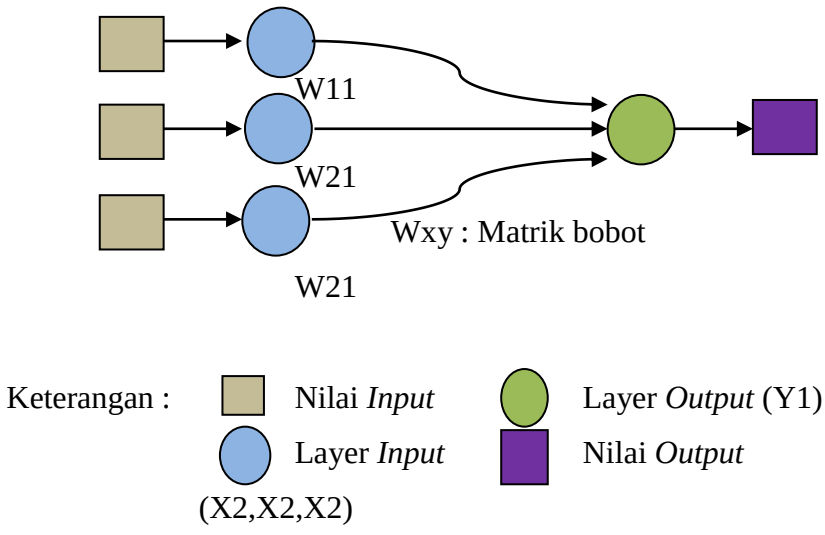
2.2.2 Jaringan Syarat Tiruan *(Artificial Neural Network/ANN)*

Jaringan Syarat Tiruan *(Artificial Neural Network/ANN)* telah banyak digunakan untuk memecahkan berbagai masalah dalam berbagai disiplin ilmu seperti bidang komputer, teknik, perdagangan dan lain-lain. ANN juga dapat digunakan untuk melakukan prediksi atau peramalan. ANN adalah Suatu neural network (NN), adalah suatu struktur pemroses informasi yang terdistribusi dan

bekerja secara paralel, yang terdir iatas elemen pemroses (yang memiliki memori lokaldan beroperasi dengan informasi lokal) yang diinter koneksi bersama dengan alur sinyal searah yang disebut koneksi (Hecht-Nielsend, 1988). Sebuah jaringan saraf adalah sebuah prosesor yang terdistribusi paralel dan mempuyai kecenderungan untuk menyimpan pengetahuan yang didapatkannya dari pengalaman dan membuatnya tetap tersedia untuk digunakan (Haykin, 1994). Sebuah ANN merupakan sistem adaptif yang merubah strukturnya berdasarkan informasi eksternal maupun internal yang mengalir melalui jaringan selama fase pembelajaran. ANN ditentukan oleh tiga hal, yaitu: pola hubungan antar neuron, metode untuk menentukan bobot penghubung dan fungsi aktivasi. Arsitektur ANN secara umum dibagi menjadi empat, yaitu *Single-Layer Feedforwad Network* (SLFN), *Multi-Layer Feedforward Network* (MLFN), *Recurrent Network* dan *Lattice Structure*.

1. Jaringan layar tunggal *(single layer network)*

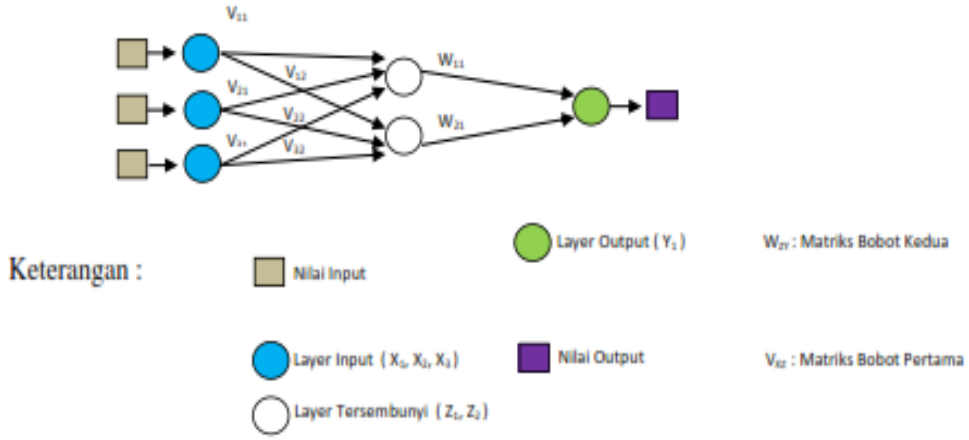
   Jaringan dengan lapisan tunggal umumnya terdiri dari 1 layar *input* dan 1 layar *output*. Setiap neuron atau unit yang terdapat pada layar *input* selalu terhubung dengan setiap neuron yang terdapat pada layar *output*. Pada gambar 2.2 ini menggambarkan arsitektur jaringan layar tunggal dengan 3 layar *input* dan 1 layar *output* dilihat pada gambar 2.2.



Gambar 2.2 Arsitektur Layar Tunggal (Siang, 2005)

2. Jaringan layar jamak *(multi layer network)*

Jaringan dengan layar jamak memiliki layar *input*, layar tersembunyi *(hidden layer)*, dan layar *output*. Jaringan dengan layar jamak mampu menyelesaikan permasalahan yang lebih kompleks dibandingkan dengan jaringan layar tunggal, namun proses pelatihan membutuhkan waktu lama dapat dilihat pada Gambar 2.3.



Gambar 2.3 Arsitektur Layar Jamak (Siang, 2005)

3. *Reccurent*

Model jaringan reccurent mirip dengan jaringan layar tunggal maupun majemuk. Hanya saja, ada *neuron output* yang memberikan sinyal pada layar atau unit *input*. Hal ini juga sering disebut dengan *feedback loop* (Siang, 2005)

Secara garis besar, *training* jaringan dengan metode perambatan balik meliputi 3 (tiga) tahap:

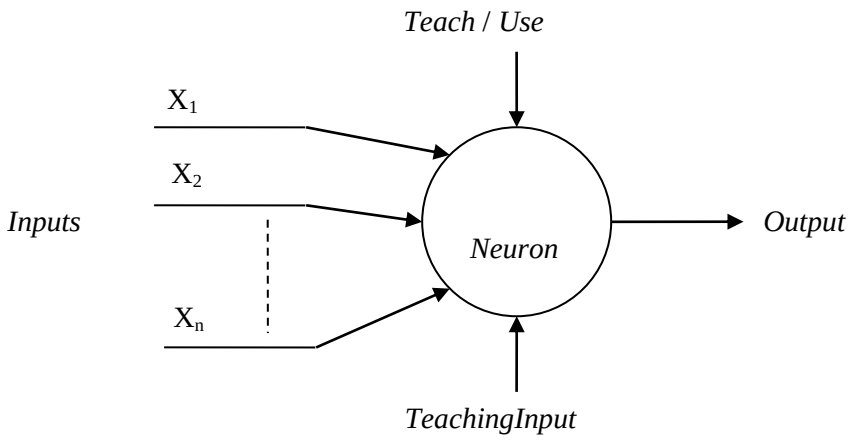
1. Tahap maju (*feedforward*)
2. Tahap perhitungan galat perambatan-balik (*backpropagation of error*)
3. Tahap pembaharuan bobot dan bias (*adjustment of the weights andbiases*).

Model arsitektur dasar dari jaringan syaraf tiruan (JST) tersebut dapat diamsusikan sebagai berikut*:*

1. Masukan sebagai masukan (xn) yang berfungsi sebagai penerima sinyal.

2. Bobot koneksi (Wjn) untuk menyimpan informasi.
3. Bias (W0) yang berfungsi mengatur daerah nilai ambang.
4. Elemen pemroses (J) dan fungsi aktifasi (F1) untuk memroses informasi.
5. Keluaran sebagai keluaran (Y1) yang akan menyampaikan hasil pemrosesan informasi ke sel berikutnya.

ANN telah banyak digunakan untuk memecahkan berbagai masalah dalam berbagai disiplin ilmu seperti bidang komputer, teknik, perdagangan dan lain-lain. ANN juga dapat digunakan untuk melakukan prediksi atau peramalan. ANN adalah sebuah model matematis atau model komputasi yang terinspirasi oleh struktur dan/atau aspek fungsional dari jaringan saraf biologi. Sebuah ANN terdiri dari sebuah kelompok yang saling berhubungan dari neuron buatan dan memproses informasi menggunakan penghubung untuk melakukan perhitungan. Sebuah ANN merupakan sistem adaptif yang merubah strukturnya berdasarkan informasi eksternal maupun internal yang mengalir melalui jaringan selama fase pembelajaran dapat dilihat pada Gambar 2.4.



Gambar 2.4. Sebuah Sel Syaraf Sederhana (Yani, 2005)

ANN ditentukan oleh tiga hal, yaitu:

a. Pola hubungan antar neuron.
b. Metode untuk menentukan bobot penghubung.

c. Fungsi aktivasi.

Umumnya, jika menggunakan ANN, hubungan antara masukan dengan keluaran harus diketahui secara pasti dan jika hubungan tersebut telah diketahui maka dapat dibuat suatu model. Ada tiga tipe pembelajaran yang dikenal dalam ANN, yaitu pembelajaran terawasi dan pembelajaran tak terawasi dan yang ketiga kombinasi dua tipe.

1. Supervisi (*supervised*)

   Dalam pelatihan dengan supervisi, terdapat sejumlah pasangan data (masukan–target keluaran) yang dipakai untuk melatih jaringan hingga didapatkan bobot jaringan yang diinginkan. Untuk setiap kali pelatihan, suatu *input* diberikan ke jaringan akan memproses dan mengeluarkan keluaran. Selisih antara keluaran jaringan dengan target merupakan kesalahan yang terjadi. Jaringan akan memodifikasi bobot sesuai dengan kesalahan tersebut (Siang, 2005). Contoh model yang di gunakan : *Perceptron, Backpropagation, ADALINE* dan *Hopfield*.

2. Tanpa Supervisi (*unsupervised*)

   Dalam jaringan kompetitif, jaringan terdiri dari dua layar atau lapisan, yaitu layar *input* dan layar kompetensi. Layar *input* menerima data eksternal. Layar kompetitif berisi neuron-neuron yang saling berkompetisi agar memperoleh kesempatan untuk merespon atau menanggapi sifat-sifat yang ada didalam data masukan. Neuron yang memenangkan kompetisi akan memperoleh sinyal yang berikutnya ia teruskan. Bobot dari neuron pemenang akan dimodifikasi sehingga menyerupai dengan data masukkan (Siang, 2005). Contoh model yang digunakan yaitu *Competitive* dan *Neocognitron*.

3. Hibrida (*hybrid*)

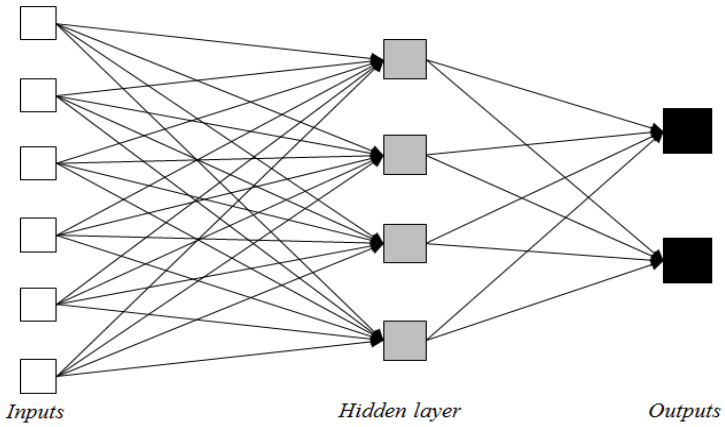
   Merupakan kombinasi dari pembelajaran supervisi dan tanpa supervisi. Sebagian dari bobot-bobotnya ditntukan melalui pembelajaran yang supervisi dan sebagiannya melalui pembelajaran tanpa supervisi. Contoh model yang digunakan yaitu algoritma RBF (Siang, 2005).

Arsitektur ANN secara umum dapat dibagi kedalam dua kategori, yaitu: Struktur Umpan Balik dan Struktur Berulang (umpan balik).

a. Struktur Umpan Balik

Sebuah jaringan yang sederhana mempunyai struktur umpan balik dimana signal bergerak dari masukan kemudian melewati lapisan tersembunyi dan akhirnya mencapai unit keluaran, yaitu mempunyai struktur perilaku yang stabil. Tipe jaringan *feed forward* mempunyai sel syaraf yang tersusun dari beberapa lapisan. Lapisan ini hanya memberikan pelayanan dengan mengenalkan suatu nilai dari suatu variabel. Lapisan tersembunyi dan lapisan *output* sel syaraf terhubung satu sama lain dengan lapisan sebelumnya.Yang termasuk dalam struktur *feed forward* dapat dilihat pada Gambar 2.5 adalah:

1. *Single-layer perceptron.*
2. *Multilayer perceptron.*
3. *Radial-basis function networks.*
4. *Higher-order networks.*
5. *Polynomial learning networks*



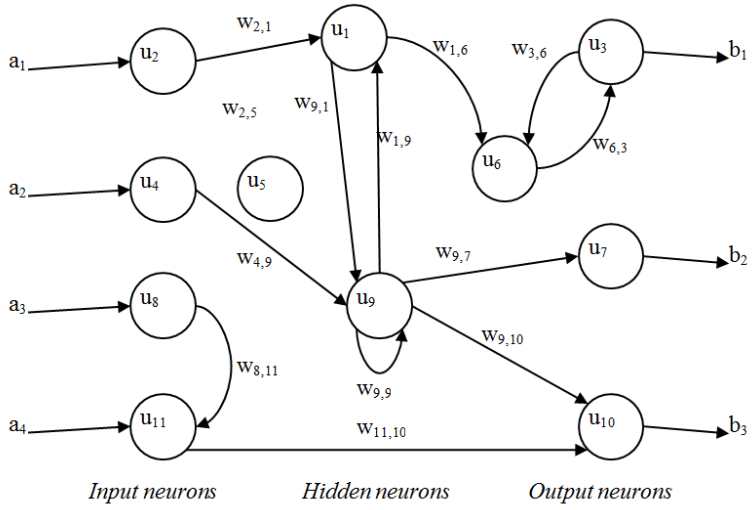
Gambar 2.5. ANN *Feedforward* (Yani, 2005)

b. Struktur Berulang (umpan balik)

Jika suatu jaringan berulang atau mempunyai koneksi kembali dari *output* ke *input*, akan menimbulkan ketidak stabilan dan akan menghasilkan dinamika yang sangat kompleks. Yang termasuk struktur *recurrent* (umpan balik) dapat dilihat pada Gambar 2.6 adalah :

1. *Competitive networks.*

2. *Self-organizing maps.*
3. *Hopfield networks.*
4. *Adaptive-resonanse theory models.*



Gambar 2.6. ANN *Feedback* (Yani, 2005)

2.2.3 RMSE *(Root Mean Square Error)*

Kriteria yang digunakan untuk mengukur kebaikan model setelah diperoleh suatu model adalah *root mean square* error (RMSE) (Wahyuningsih, 2012). RMSE merupakan alat seleksi model berdasarkan pada *error* hasil estimasi. *Error* yang ada menunjukkan seberapa besar perbedaan hasil estimasi dengan nilai yang akan diestimasi. Nilai ini akan digunakan untuk menentukan model mana yang terbaik. Definisi RMSE dapat ditulis sebagai berikut.

$$\text{RMSE} = \sqrt{\frac{\sum_{i}^{n=1} (yi\ Yi)2}{n}} \qquad (2.2)$$

RMSE : *Root Mean Square Error*

n : Jumlah Sampel

yi : Nilai Aktual

Yi : Nilai Prediksi